IMS Collections
Contemporary Developments in Bayesian Analysis and Statistical Decision Theory: A Festschrift for William E. Strawderman
Vol. 8 (2012) 1–23

DOI: 10.1214/11-IMSCOLL801

Density estimation in high and ultra high dimensions, regularization, and the L_1 asymptotics

Anirban DasGupta[1,*] **and S.N. Lahiri**[2]

Purdue University and Texas A&M University

Abstract: This article gives a theoretical treatment of the asymptotics of the L_1 error of a model-based estimate of a density $f(x|\theta)$ on a finite dimensional Euclidean space $\mathcal{R}^k$.

The dimension p of the parameter vector θ is considered arbitrary but fixed in Section 2. Two theorems in Section 2 lay out the weak limits of a suitably scaled L_1 error, with a general estimating sequence $\hat{\theta}$ and a general family of smooth densities $f(x|\theta)$ dominated by some σ-finite measure, the discrete case included. We show that the L_1 error converges at the coarsest rate corresponding to the different coordinates of the parameter vector θ. Four applications are detailed in Section 3, a special one being a new confidence interval for a Poisson mean.

Section 4 considers the high and the ultra high dimensional case, where p grows with n. The exact critical growth rate for p when maximum likelihood starts to falter is derived. Maximum likelihood is shown to exhibit a trichotomy of behavior; the desired behavior below the threshold, problematic behavior at the threshold, and disastrous behavior above the threshold.

It is then shown that regularization, if coupled with the right amount of sparsity, can return consistent density estimation, even at the best possible $n^{-1/2}$ rate. We give a complete description of the limiting behavior of the regularized density estimate under different sparsity conditions. Section 4 is specialized to the Gaussian case due to its special importance and well known links to function estimation.

Contents

*Research partially supported by NSF grant no. DMS 0707139.
[1]150 N. University Street West Lafayette, IN 47907-2068, e-mail: dasgupta@stat.purdue.edu
[2]Department of Statistics, 525J Blocker, TAMU-3143 Texas A & M University, e-mail: snlahiri@stat.tamu.edu
AMS 2000 subject classifications: Primary 60K35, 60K35; secondary 60K35
Keywords and phrases: density estimation, L_1 error, consistency, asymptotics, maximum likelihood, regularization, James-Stein estimator, sparsity, Gaussian

1. Introduction

Inspired by practical problems in several subject matter fields, a large amount of theoretical work is currently being done on the so called *increasing dimension* inferential problems. Depending on the exact context, the number of parameters p is assumed to increase at a suitable rate with the number of observations n in these problems. Maximum likelihood, which is asymptotically unbeatable in the fixed dimension case under enough conditions, fails when p increases too rapidly with n. To have any hope of accurate inference in such problems, even asymptotically, one must *regularize*. The exact form of regularization, once again, depends on the particular context; but the common idea is that one must use external information about the unknown parameters and employ alternative procedures that use the external information. If the information is correct, the regularized procedure works better than ordinary maximum likelihood. For instance, for the seemingly innocuous problem of estimating a normal mean vector, maximum likelihood will be no good at all if p is large. External information in such a problem may be that a large number of the true means are zero, or close to zero, often known as *sparsity* of the mean vector. An estimate which can exploit the sparsity will perform better than ordinary maximum likelihood. Such an estimate is called a regularized estimate.

The purpose of this paper is to study estimation of a parametric density in high or ultra high dimensions, and to pin down exactly when maximum likelihood fails and how suitable regularization, coupled with just the right amount of sparsity, can save the situation. See Bickel and Li (2006) for a modern overview of the current state of the area and Liu, Lafferty, and Wasserman (2007) for a specific proposal for density estimation in such high dimensional cases. The word *dimension* refers to the affine dimension of the parameter space. As we remarked above, such problems with far too many parameters and relatively less samples have become important in several areas of application of statistics, and are also theoretically important due to their demonstrated connections to various nonparametric problems, as in Ibragimov and Has'minskii (1977), Nussbaum (1996), Brown and Low (1996), Donoho and Johnstone (1998), Johnstone (2003), and Cai and Low (2005), among numerous others. Precise additional theoretical development is clearly needed to catch up with the procession of methodologies.

As our criterion, we use the L_1 error

$$\int_{\mathcal{R}^k} |\hat{f}_n(x) - f(x)| d\mu(x). \tag{1.1}$$

We chose the L_1 error for several reasons, the primary being that it allows us to make statements simultaneously about estimating probabilities of arbitrary measurable sets, and its well known invariance properties with respect to transformations and the dominating measure. A standard reference is Devroye and Györfi (1984). Of course, we may also use L_p errors for other values of p; a standard reference is Hall (1984). Actually, the L_1 error is technically more difficult to handle than the Hellinger metric, which is topologically equivalent to the L_1 error. The Hellinger metric in the context of the problems of this paper will be treated in a future article. An interesting fact which is not obvious, but can be proved in this particular problem, is that consistency with respect to the L_1 error is equivalent to consistency with respect to any of the Kolmogorov metric, the L_2 distance, the Hellinger metric, and the Kullback-Leibler distance.

Fully nonparametric density estimation is problematic in as few as five dimensions. Against this background, perhaps it is not surprising that parametrics, and especially Gaussian parametrics, have made a dramatic come back in these modern high dimensional problems. We follow that tradition while dealing with density estimation in high dimensions. Thus, the true underlying density f is taken to be a Radon-Nikodym derivative with respect to a suitable σ-finite measure on some $\mathcal{R}^k, 1 \leq k < \infty$, and is assumed to be indexed by a p-dimensional parameter. As is usual, we assume that we have n iid observations $X_1, X_2, \cdots, X_n$ to estimate the true f. All results in the paper are in an asymptotic paradigm, i.e., as $n \to \infty$.

As regards p, the dimension of the parameter space, the results are clearly divided into two different setups. Section 2 details our results on consistency and weak limits of the L_1 error when p is arbitrary but fixed. In other words, p does not grow with n in Section 2. The densities are general smooth densities, and the assumptions required are fairly minor. In Section 3, the theorems of Section 2 are applied to four concrete examples. One example uses a theorem in Section 2 to propose a new confidence interval for a Poisson mean. Le Cam (1990) had previously suggested construction of confidence sets by using metrics on probability measures. Two other examples lay out the general location-scale parameter problem as a special example of the general theorem. A fourth example applies a general theorem in Section 2 to robust estimation of a p-dimensional location parameter and demonstrates a striking robustness property of the L_1 median. Section 2 also gives a theorem for the two-sample case, for which we use as our criterion the corresponding L_1 error

$$\int_{\mathcal{R}^k} |\hat{f}_{1,n}(x) - \hat{f}_{2,m}(x)| d\mu(x), \tag{1.2}$$

where $\hat{f}_{1,n}(x), \hat{f}_{2,m}(x)$ are two different estimate sequences of a common true underlying $f(x)$. Usually, this common f will come from some sort of a null hypothesis that two populations have an underlying common distribution.

In Section 4, we make our transition to the *high and ultra high* dimensions. In other words, now p grows with n. We make the restriction to a Gaussian density in this section. Clearly, the Gaussian case should be the first case to try, although some of the results in Section 4 should admit easy generalizations. We show that $p \sim n$ is when maximum likelihood starts to fail. If $p \sim n$, maximum likelihood fails to deliver even consistent density estimates. If $p = o(n)$, then maximum likelihood works and it converges at the $n^{-\frac{1}{2}}$ rate. If p grows faster than n, then we show that maximum likelihood not only fails, but fails miserably.

We then show that regularization can save us when p grows at the rate of n, or even if it grows faster than n. However, the faster it grows, the more sparse will have to be the underlying Gaussian mean vector. As a sample of one such result, we show that at whatever rate p grows, if $\frac{n}{p}(\theta'\theta) \to 0$ and $n(\theta'\theta) \to b \geq 0$ (but $b < \infty$), the regularized density estimate will still succeed in not only delivering on consistency, but in also producing the best possible $n^{-\frac{1}{2}}$ rate.

We have regularized the MLE by using the James-Stein estimate. The sparsity conditions corresponding to this specific regularization work out to conditions on the L_2 norm of the mean vector θ. It is certainly desirable to use other methods of regularization, e.g., hard thresholding. If we do, the sparsity conditions will work out to something else. This is not explored here, and is currently under investigation.

2. Asymptotics for fixed dimensions

The basic problem of this section is the following: $f(x|\theta)$ is a density with respect to some dominating measure μ on a finite dimensional Euclidean space $\mathcal{R}^k$. The parameter θ is a p-dimensional Euclidean vector for some fixed but arbitrary $p < \infty$. Based on an iid sample $X_1, X_2, \cdots, X_n$ from $f(x|\theta)$, we first estimate θ by using some fairly general estimator sequence $\hat{\theta} = \hat{\theta}(X_1, X_2, \cdots, X_n)$. We then look at the implied model-based density estimate $f(x|\hat{\theta})$, and study consistency and weak convergence of the L_1 error $\int_{\mathcal{R}^k} |f(x|\hat{\theta}) - f(x|\theta)| d\mu(x)$. The results are then used in four concrete applications. The two sample case is also supplied.

2.1. The theorems

Let $P_\theta, \theta \in \Theta \subset \mathcal{R}^p$ be a parametric family of distributions on $(\mathcal{R}^k, \mathcal{B}(\mathcal{R}^k))$ such that $P_\theta \ll \mu$ for some σ-finite measure μ on $(\mathcal{R}^k, \mathcal{B}(\mathcal{R}^k))$, where $\mathcal{B}(\mathcal{R}^k)$ denotes the Borel σ-field on $\mathcal{R}^k$, $1 \le k < \infty$. Let $f_\theta \equiv \frac{dP_\theta}{d\mu}$. Let $\mathcal{Z}_+ = \{0, 1, \ldots\}$ denote the set of all nonnegative integers. For a matrix Γ, let Γ' denote its transpose. We define the derivative

$$D^\alpha f_\theta(x) = \frac{\partial^\alpha}{\partial \theta_1^{\alpha_1} \cdots \partial \theta_p^{\alpha_p}} f_\theta(x),$$

where $\alpha = (\alpha_1, \ldots, \alpha_p) \in \mathcal{Z}_+^p$ with $|\alpha| = \sum_{j=1}^p \alpha_j$. Let $\hat{\theta}_n$ be an estimator of θ and let r be a given positive integer such that at the true value $\theta = \theta_0$, the following conditions hold:

(A.1) There exist a nondegenerate random vector Z and a sequence of constants $c_n \to \infty$ such that

$$c_n(\widehat{\theta}_n - \theta_0) \stackrel{\mathcal{L}}{\Rightarrow} Z \text{ as } n \to \infty.$$

(A.2)$_r$ There exists a set $A \in \mathcal{B}(\mathbb{R}^k)$ with $\mu(A^c) = 0$ such that, for all $x \in A$, $f_\theta(x)$ is $(r+1)$-times continuously differentiable in some neighborhood of θ_0, and for all $\alpha \in \mathcal{Z}_+^p$ such that $1 \le |\alpha| \le r$,

$$\int_A \|D^\alpha f_{\theta_0}(x)\| \mu(dx) < \infty \tag{2.1}$$

$$\int_A \sup_{\|t-\theta_0\| \le \delta,\, |\alpha| = r+1} \|D^\alpha f_t(x)\| \mu(dx) < \infty \qquad \text{for some } \delta > 0. \tag{2.2}$$

Conditions (A.1) and (A.2)$_r$ are satisfied in many applications. In this section, we shall use condition (A.2)$_r$ with $r = 1$ only. We verify these conditions in some examples in the next Section.

Under the above conditions, the first theorem below gives the limiting distribution of the L_1 error for the plug-in parametric density estimator.

Theorem 2.1. (The One Sample Case). *Suppose that conditions (A.1) and (A.2)$_r$ hold with $r = 1$. Then*

$$c_n \|f_{\widehat{\theta}_n} - f_{\theta_0}\|_1 \stackrel{\mathcal{L}}{\Rightarrow} \int |Z' f_{\theta_0}^{(1)}(x)| \mu(dx), \tag{2.3}$$

where the random integral on the right is defined in a pointwise sense and $f_\theta^{(1)}(x)$ denotes the vector of first order partial derivatives of $f_\theta(x)$ w.r.t θ.

Proof. In $(A.2)_1$, for notational simplicity, set $A^c = \emptyset$. Then,

$$\begin{aligned} & \|f_{\widehat{\theta}_n} - f_{\theta_0}\|_1 \\ = & \int |f_{\hat{\theta}_n} - f_{\theta_0}| d\mu \\ = & \int |\{f_{\theta_0}(x) + (\hat{\theta}_n - \theta_0)' f^{(1)}_{\theta_0}(x) + R_n(x)\} - f_{\theta_0}(x)| \mu(dx) \\ = & \int |(\hat{\theta}_n - \theta_0)' f^{(1)}_{\theta_0}(x) + R_n(x)| \mu(dx) \end{aligned} \tag{2.4}$$

$$\begin{cases} \geq & \int |(\hat{\theta}_n - \theta_0)' f^{(1)}_{\theta_0}(x)| \mu(dx) - \int |R_n(x)| \mu(dx) \\ \leq & \int |(\hat{\theta}_n - \theta_0)' f^{(1)}_{\theta_0}(x)| \mu(dx) + \int |R_n(x)| \mu(dx), \end{cases} \tag{2.5}$$

where $R_n(x)$ denotes the remainder term. By (A.1) and $(A.2)_1$,

$$\begin{aligned} & \int |R_n(x)| \mu(dx) \\ \leq & \text{ const. } \| \hat{\theta}_n - \theta_0 \|^2 \int \sup_{\|t-\theta_0\| \leq \|\hat{\theta}_n - \theta_0\|, |\alpha|=2} |D^\alpha f_t(x)| \mu(dx) \\ = & O_p(c_n^{-2}). \end{aligned} \tag{2.6}$$

Next, note that the mapping

$$h(t) \equiv \int |t' f^{(1)}_{\theta_0}(x)| \mu(dx), \quad t \in \mathcal{R}^p$$

is continuous on $\mathcal{R}^p$. This follows from the DCT (the dominated convergence theorem):

$$\begin{aligned} h(t+s) \equiv & \int |(t+s)' f^{(1)}_{\theta_0}| \mu(dx) \\ \to & \int |t' f^{(1)}_{\theta_0}(x)| \mu(dx) \text{ as } \| s \| \to 0. \end{aligned}$$

Hence, by (7) and (8), (5) follows. □

The corresponding theorem for the two-sample case is the following. Theorem 2.2 is useful for testing on the basis of independent samples from F_θ and F_λ that $\theta = \lambda$. The hypothesis can be rejected if $\|f_{\hat{\theta}_{1,n}} - f_{\hat{\theta}_{2,n}}\|_1$ is large, with the cut-off value coming from the limiting null distribution given below in Theorem 2.2.

Theorem 2.2. (The Two Sample Case). *Suppose that the conditions of Theorem 2.1 hold.*

(a) Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} P_{\theta_0}$, all observations being mutually independent. Let $\hat{\theta}_{1,n} = \hat{\theta}_{1,n}(X_1, X_2, \ldots, X_n)$, and $\hat{\theta}_{2,n} = \hat{\theta}_{2,n}(Y_1, Y_2, \ldots, Y_n)$ be two estimators, each satisfying the conditions in Theorem 2.1 with the same sequence $\{c_n\}$. Then,

$$c_n \|f_{\hat{\theta}_{1,n}} - f_{\hat{\theta}_{2,n}}\|_1 \overset{\mathcal{L}}{\Rightarrow} \int |(Z_1 - Z_2)' f^{(1)}_{\theta_0}(x)| \mu(dx), \tag{2.7}$$

where Z_1, Z_2 are independent, being as in assumption (A.1).

(b) Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} P_{\theta_0}$, all observations being mutually independent. Let $\hat{\theta}_{1,m} = \hat{\theta}_{1,m}(X_1, X_2, \ldots, X_m)$, and $\hat{\theta}_{2,n} = \hat{\theta}_{2,n}(Y_1, Y_2, \ldots, Y_n)$ each satisfy the conditions in Theorem 2.1 with the the same sequence $\{c_n\}$. Assume further that there exists a function $c(.)$ of regular variation at ∞ with exponent $\gamma > 0$ such that $c_n = c(n)$, and that $m, n \to \infty$ in a way that $\frac{m}{m+n} \to \tau$ for some $0 \le \tau \le 1$. Then,

$$\frac{c_m c_n}{c_{m+n}} \| f_{\hat{\theta}_{1,m}} - f_{\hat{\theta}_{2,n}} \|_1 \overset{\mathcal{L}}{\Rightarrow} \int |[(1-\tau)^\gamma Z_1 - \tau^\gamma Z_2]' f^{(1)}_{\theta_0}(x)| \mu(dx), \tag{2.8}$$

where Z_1, Z_2 are independent, and as in assumption (A.1).

Proof. We only outline the proof of part (b). First note that $\frac{c_n}{c_{m+n}} = \frac{c(\frac{n}{m+n}(m+n))}{c(m+n)}$. Since $\frac{n}{m+n} \to 1-\tau$ and $c(.)$ is of regular variation at ∞ with exponent γ, by the local uniformity of the regular variation property (see pp. 17, Resnick (1987)), $\frac{c_n}{c_{m+n}} \to (1-\tau)^\gamma$. Similarly, $\frac{c_m}{c_{m+n}} \to \tau^\gamma$. Part (b) of Theorem 2.2 now follows on writing $\frac{c_m c_n}{c_{m+n}} \| f_{\hat{\theta}_{1,m}} - f_{\hat{\theta}_{2,n}} \|_1$ as $\int |\frac{c_n}{c_{m+n}} c_m (f_{\hat{\theta}_{1,m}} - f_{\theta_0}) - \frac{c_m}{c_{m+n}} c_n (f_{\hat{\theta}_{2,n}} - f_{\theta_0})| \mu(dx)$, and by using the Taylor series argument exactly as in Theorem 2.1. □

3. Examples

Next we consider a number of examples that illustrate Theorem 2.1.

Example 1 (A New Poisson Confidence Interval). Let $P_\theta = \text{POISSON}(\theta)$, $\theta \in (0, \infty)$, and let $\widehat{\theta}_n = \bar{X}_n$. The standard estimate of θ is indeed the MLE $\overline{X}_n$. If we estimate P_θ by the Poisson distribution with mean $\overline{X}$, then the total variation distance between P_θ and $P_{\overline{X}}$ is related to the L_1 error by the expression

$$d_{TV}(POI(\theta), POI(\overline{X})) = \frac{1}{2} \| f_\theta - f_{\overline{X}} \|_1.$$

We work out the limiting distribution of this total variation distance in this example. Possible practical applications of this limiting distribution are indicated at the end of this example.

By the central limit theorem,

$$\sqrt{n}(\widehat{\theta}_n - \theta) \overset{\mathcal{L}}{\Rightarrow} N(0, \theta)$$

for all θ and hence, (A.1) holds. Next note that (writing D for $\frac{\partial}{\partial \theta}$),

$$f_\theta(x) \equiv e^{-\theta} \frac{\theta^x}{x!}, \quad x = 0, 1, \ldots$$

$$\Rightarrow Df_\theta(x) = \begin{cases} e^{-\theta} \left(\frac{\theta^{x-1}}{(x-1)!} - \frac{\theta^x}{x!} \right), & \text{if } x = 1, \ldots \\ -e^{-\theta} & \text{if } x = 0, \end{cases}$$

$$D^2 f_\theta(x) = \begin{cases} e^{-\theta} \left[\left(\frac{\theta^{x-2}}{(x-2)!} - \frac{\theta^{x-1}}{(x-1)!} \right) - \left(\frac{\theta^{x-1}}{(x-1)!} - \frac{\theta^x}{x!} \right) \right], & x = 2, 3 \ldots \\ e^{-\theta}[-1 - (1-\theta)], & x = 1 \\ e^{-\theta}, & x = 0 \end{cases}$$

$$= \begin{cases} e^{-\theta} \left[\frac{\theta^x}{x!} - \frac{2\theta^{x-1}}{(x-1)!} + \frac{\theta^{x-2}}{(x-2)!} \right], & x = 2, 3 \ldots \\ e^{-\theta}(\theta - 2), & x = 1 \\ e^{-\theta}, & x = 0. \end{cases}$$

Hence, it follows that

$$\begin{aligned} & \sup_{|t-\theta|\le\delta} |D^2 f_t(x)| \\ \le\ & e^{-(\theta-\delta)} \left[\frac{(\theta+\delta)^2}{x!} + \frac{(\theta+\delta)^{x-2}}{(x-2)!} I(x=2,\ldots) \right], \end{aligned}$$

which, in turn, implies that

$$\begin{aligned} & \int \sup_{|t-\theta|\le\delta} |D^2 f_t(x)| \mu(dx) \\ \le\ & e^{-(\theta-\delta)} \left[\sum_{x=0}^{\infty} \frac{(\theta+\delta)^x}{x!} + \sum_{x=2}^{\infty} \frac{(\theta+\delta)^{x-2}}{(x-2)!} \right] \\ =\ & e^{-(\theta-\delta)} [2e^{(\theta+\delta)}] \\ =\ & 2e^{2\delta} < \infty. \end{aligned}$$

Thus, the second assumption in $(A.2)_1$ holds.

Also,

$$\begin{aligned} \tau_1(\theta) &\overset{def}{\equiv} \int |Df_\theta(x)| \mu(dx) \\ &= e^{-\theta} + \sum_{x=1}^{\infty} \left| e^{-\theta} \left(\frac{\theta^{x-1}}{(x-1)!} - \frac{\theta^x}{x!} \right) \right| \\ &= e^{-\theta} + \sum_{x=1}^{\infty} e^{-\theta} \frac{\theta^x}{x!} \left| \frac{x}{\theta} - 1 \right| \\ &= \sum_{x=0}^{\infty} e^{-\theta} \frac{\theta^x}{x!} \left| \frac{x}{\theta} - 1 \right| \\ &= E_\theta \left| \frac{X_1}{\theta} - 1 \right|, \end{aligned}$$

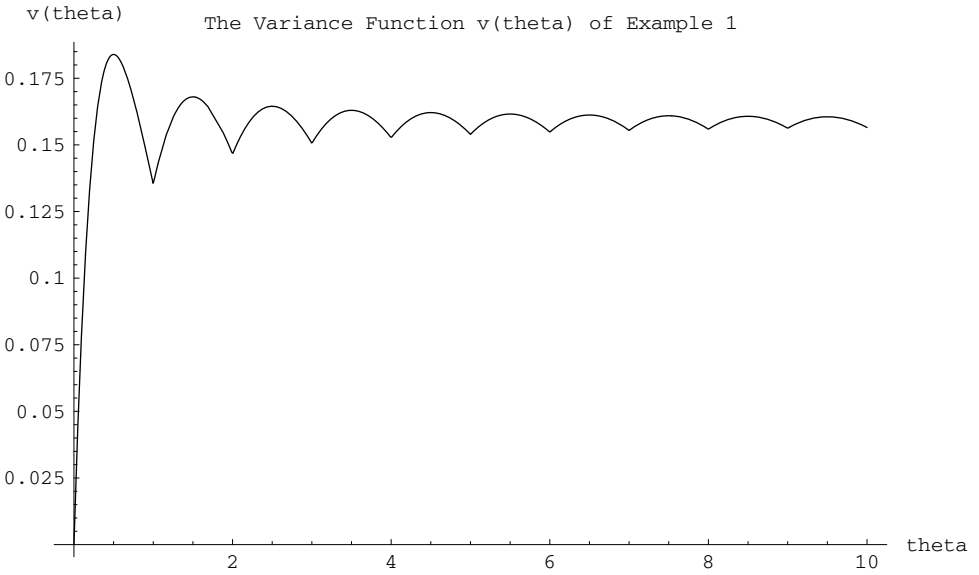
which implies the first assumption in $(A.2)_1$. Hence,

$$\sqrt{n} d_{TV}\Big(POI(\bar{X}_n), POI(\theta) \Big) \overset{\mathcal{L}}{\Rightarrow} |N(0,\theta)| \cdot \frac{\tau_1(\theta)}{2} \quad \text{for all} \quad \theta \in (0,\infty).$$

Interestingly, there is an exact expression for $\tau_1(\theta)$. Indeed, $\tau_1(\theta) = \frac{2e^{-\theta}\theta^{[\theta]}}{[\theta]!}$; see Diaconis and Zabell (1991). Here, $[\theta]$ denotes the integer part of θ. Thus, $\sqrt{n} d_{TV}(POI(\bar{X}), POI(\theta))$ converges to the absolute value of a normal random variable with mean 0 and variance

$$v(\theta) = \frac{e^{-2\theta}\theta^{2[\theta]+1}}{([\theta]!)^2}. \tag{3.1}$$

The function $v(\theta)$ has an interesting shape and is plotted below. We notice from the plot that as distributions, $POI(\bar{X})$ and $POI(\theta)$ are the closest together when θ is an integer, and the farthest apart when θ is a half-integer. It would be interesting to give an intuitive explanation for this finding.



Discussion. What is a possible application of this result? As mentioned in the introduction, the result can be used to construct an asymptotically correct confidence set for a Poisson mean θ. There are a few possibilities here. The most obvious one is to construct the confidence set as

$$\Big\{\theta : \sqrt{n} d_{TV}\Big(POI(\overline{X}), POI(\theta)\Big) \le \sqrt{v(\overline{X})}\chi_\alpha\Big\}, \tag{3.2}$$

where $v(.)$ is the function described above and χ^2_α is the $(1-\alpha)$th percentile of a $\chi^2(1)$ distribution. Another possibility is to exploit the connection of $v(\theta)$ to the mean absolute deviation function $E(|X_1 - \theta|)$ and directly estimate the mean absolute deviation by the sample mean absolute deviation, namely $\frac{1}{n}\sum_{i=1}^n |X_i - \overline{X}|$. This will also result in an asymptotically correct confidence set. The *Wald confidence interval* $\overline{X} \pm z_{\alpha/2}\sqrt{\frac{\overline{X}}{n}}$ is a textbook confidence set for θ, which has been shown to have poor coverage properties in Brown, Cai, and DasGupta (2003). It would be interesting to further pursue these two confidence intervals for θ and study their coverage properties.

Example 2 (The General Location-Scale Family). Let

$$F_\theta(x) = F_0\left(\frac{x-\mu}{\sigma}\right), \quad \theta = (\mu, \sigma)' \in \mathbb{R} \times (0, \infty).$$

Suppose that $\widehat{\theta}_n$ is some sequence of estimates such that

(A.1)′ $\sqrt{n}(\widehat{\theta}_n - \theta) \overset{\mathcal{L}}{\Rightarrow} N(0, \Sigma_0)$, and

(A.2)′$_1$ for some $\delta > 0$ and for $i = 0, 1, 2$,

$$\int \sup_{\|\theta - t\| \le \delta} \left| \left[\frac{1}{t_2} f_0^{(i)}\left(\frac{x - t_1}{t_2}\right)\right]\right| \mu(dx) < \infty,$$

where $t = (t_1, t_2)'$ and where $f_0^{(i)}$ denotes the ith derivative of f_0 for $i \ge 1$ and where $f_0^{(0)} = f_0$.

A word of notational caution is that the location parameter and the dominating measure have both been denoted as μ, to preserve the consistency of notation for the dominating measure.

Now,

$$\begin{aligned} f_\theta(x) &= \frac{1}{\sigma} f_0\left(\frac{x-\mu}{\sigma}\right) \\ &\Rightarrow \begin{cases} \frac{d}{d\mu} f_\theta(x) = -\frac{1}{\sigma^2} f_0^{(1)}\left(\frac{x-\mu}{\sigma}\right) \\ \frac{d}{d\sigma} f_\theta(x) = \frac{-1}{\sigma^2} f_0\left(\frac{x-\mu}{\sigma}\right) + \frac{\mu}{\sigma^3} f_0^{(1)}\left(\frac{x-\mu}{\sigma}\right) \end{cases} \\ &\Rightarrow \begin{cases} \frac{d^2}{d\mu^2} f_\theta(x) = \frac{1}{\sigma^3} f_0^{(1)}\left(\frac{x-\mu}{\sigma}\right) \\ \frac{d^2}{d\sigma d\mu} f_\theta(x) = \frac{2}{\sigma^3} f_0^{(1)}\left(\frac{x-\mu}{\sigma}\right) - \frac{\mu}{\sigma^4} f_0^{(2)}\left(\frac{x-\mu}{\sigma}\right) \\ \frac{d^2}{d\sigma^2} f_\theta(x) = \left[\frac{2}{\sigma^3} f_0\left(\frac{x-\mu}{\sigma}\right) - \frac{\mu}{\sigma^4} f_0^{(1)}\left(\frac{x-\mu}{\sigma}\right)\right] \\ \qquad\qquad - \frac{3\mu}{\sigma^4} f_0^{(1)}\left(\frac{x-\mu}{\sigma}\right) + \frac{\mu^2}{\sigma^5} f_0^{(2)}\left(\frac{x-\mu}{\sigma}\right). \end{cases} \end{aligned}$$

Thus (2.4) of $(A.2)_1$ follows from $(A.2)'_1$. Next, note that

$$\begin{aligned} \sum_{j=1}^{2} \int |D_j f_\theta(x)| \mu(dx) &\leq \left[\int \frac{1}{\sigma^2} |f_0^{(1)}\left(\frac{x-\mu}{\sigma}\right)| \mu(dx)\right] \\ &+ \left[\int \frac{1}{\sigma^2} f_0\left(\frac{x-\mu}{\sigma}\right) \mu(dx) + |\frac{\mu}{\sigma^3}| \int |f_0^{(1)}\left(\frac{x-\mu}{\sigma}\right)| \mu(dx)\right] \\ &= \frac{1}{\sigma} \int |f_0^{(1)}| \mu(dx) + \left[\frac{1}{\sigma} + \frac{|\mu|}{\sigma^2} \int |f_0^{(1)}(x)| \mu(dx)\right] \\ &< \infty, \text{by } (A.2)'. \end{aligned}$$

Thus, (2.3) of $(A.2)_1$ holds. Hence, by Theorem 2.1,

$$\begin{aligned} &\sqrt{n} \| f_{\widehat{\theta}_n} - f_\theta \|_1 \\ &\overset{\mathcal{L}}{\Rightarrow} \int \left| Z_1 \left(-\frac{1}{\sigma^2} f_0^{(1)}\left(\frac{x-\mu}{\sigma}\right)\right) + Z_2 \left(-\frac{1}{\sigma^2} f_0\left(\frac{x-\mu}{\sigma}\right) + \frac{\mu}{\sigma^3} f_0^{(1)}\left(\frac{x-\mu}{\sigma}\right)\right)\right| \mu(dx) \\ &= \int \left| \frac{Z_1}{\sigma} f_0^{(1)}(x) + Z_2 \left(\frac{1}{\sigma} f_0(x) - \frac{\mu}{\sigma^2} f_0^{(1)}(x)\right)\right| \mu(dx), \end{aligned}$$

where $Z = (Z_1, Z_2)' \sim N(0, \Sigma_0)$. Notice that the limiting random variable is the L_1 - norm (wrt μ) of a Gaussian process. In general, writing the distribution of the L_1-norm of a Gaussian process is hard. For special choices of the density f_0, we can pin it down. We do so in a number of further examples below.

Example 3 (The Gaussian Case in Explicit Form). This example is a natural illustration of our general result in Example 2. Let $X_1, X_2, \ldots, X_n$ be iid from $N(\mu, \sigma^2)$, and as estimates of μ, σ, consider the usual estimates $\overline{X}, s$. The limiting distribution of $\sqrt{n} d_{TV}(P_\theta, P_{\widehat{\theta}_n})$ will be worked out in this example. Note that this distribution is independent of the true values of μ, σ due to the equivariant nature of the estimates $\overline{X}, s$. We thus set $\mu = 0, \sigma = 1$ in applying the general result enunciated in Example 2.

By a direct calculation, the integrand in the result of Example 2 works out to:

$$Z_1 f_0^{(1)}(x) + Z_2 f_0(x) \quad = \quad \frac{1}{\sqrt{2\pi}} e^{-x^2/2} |Z_2 - Z_1 x|. \tag{3.3}$$

This can be integrated in closed form, resulting in the fact that in this example, $\sqrt{n} d_{TV}(P_\theta, P_{\widehat{\theta}_n})$ converges in distribution to the distribution of $\frac{|Z_1|}{2}[2C\Phi(C) + 2\phi(C) - C]$, where $C = \frac{Z_2}{Z_1}$, Z_1, Z_2 being as in Example 2 (i.e., Z_1, Z_2 are independent normals with means zero and variances 1 and 2 respectively.

Fortunately, we can make further analytical progress. The reason is that the function $H(C) = 2C\Phi(C) + 2\phi(C) - C$ is an even function of C and monotone increasing for $C > 0$. As a consequence, the CDF of our limiting distribution, namely,

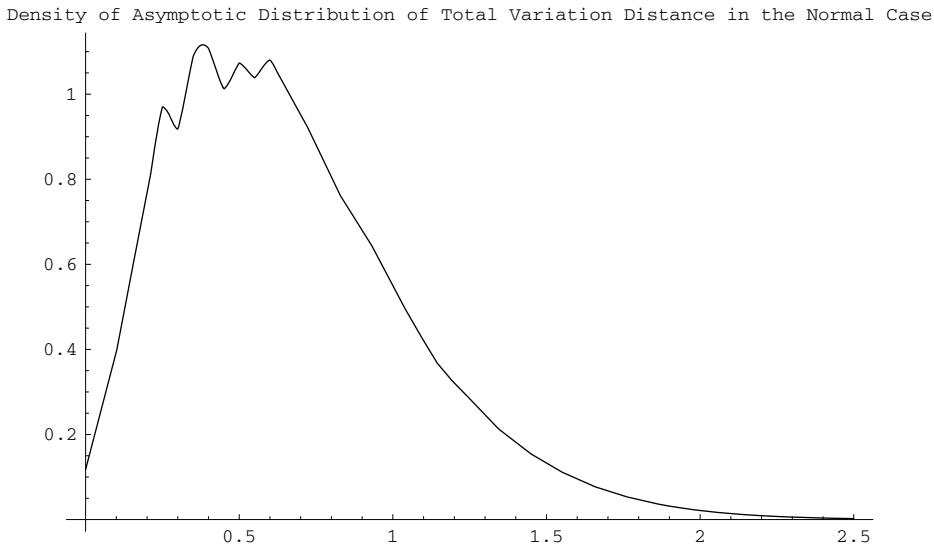
$$P(\frac{|Z_1|}{2} H(C) \le x) = P(|C| \le H^{-1}(\frac{2x}{|Z_1|}))$$

$$= E[P(|C| \le H^{-1}(\frac{2x}{|Z_1|}))|Z_1] = 2E[\Phi(\frac{|Z_1|}{\sqrt{2}} H^{-1}(\frac{2x}{|Z_1|}))] - 1,$$

on using the fact that the conditional CDF of $|C|$ given $Z_1, P(|C| \le a|Z_1 = z) = 2\Phi(\frac{a|z|}{\sqrt{2}}) - 1$, which can be established by a simple calculation. On differentiation, the density function of our limiting CDF equals the expression

$$2\sqrt{2} \int_{-\infty}^{\infty} \frac{\phi(zH^{-1}(\frac{2x}{|z|})/\sqrt{2})\phi(z)}{H'(H^{-1}(\frac{2x}{|z|}))} dz, \tag{3.4}$$

where H' denotes the derivative of H. Thus, the density of our limiting CDF can in fact be written as a one dimensional integral, which is easy to compute and plot. We computed this density at a finite grid of arguments and then smoothed it by using a default smoother on *Mathematica*. We provide a (smoothed) plot of this limiting density below. It is not certain that the bumps in the plot are real, because the plot is a smoothed version of a discrete set of values.



Example 4 (Application to Robust Estimation). This example illustrates the utility of our approach in evaluating the robustness of estimators. It also shows a rather remarkable property of the L_1 median as an estimator of a multidimensional location parameter. For instance, the property is not shared by the sample mean vector.

Let $X_1, X_2, \ldots, X_n$ be iid observations from a location parameter density $f(x-\theta)$ in $\mathcal{R}^p$. We consider the spherically symmetric case for illustration. Thus, the null density $f(x) = h(r)$, for some function $h(r)$, where $r = ||x||$. It is assumed that h is once differentiable in order that Theorem 2.1 can be applied. We will compare the limiting distributions of $\sqrt{n}d_{TV}(P_\theta, P_{\hat{\theta}_n})$ for two choices of $\hat{\theta}_n$, namely $\hat{\theta}_n = \overline{\mathbf{X}}$, and $\hat{\theta}_n = \mathbf{M_n}$, the L_1 median. The L_1 median is chosen as a specific illustration. See Brown (1983), and Small (1990) for various properties of the L_1 median.

It is well known that for a general location parameter distribution in p-dimensions, $\sqrt{n}(\mathbf{M_n} - \theta) \stackrel{\mathcal{L}}{\Rightarrow} N(\mathbf{0}, B^{-1}AB^{-1})$, where $A = E_F \frac{XX^T}{||X||^2}$, and $B = E_F \frac{I - \frac{XX^T}{||X||^2}}{||X||}$, where X denotes a single observation from the null distribution, here denoted as F. See Brown (1983), Hettmansperger and McKean (1998). Specialized to the spherically symmetric case, the expressions for A, B simplify to $A = \frac{1}{p}I$, and $B = \frac{p-1}{p}(E_F \frac{1}{||X||})I$. Thus, in the notation of Theorem 2.1, $Z \sim N(\mathbf{0}, \frac{p}{(p-1)^2(E_F \frac{1}{||X||})^2} I)$. According to Theorem 2.1, we need to find the limiting distribution of $\frac{1}{2}\int |Z' f_0^{(1)}(x)|dx$, where $f_0^{(1)}(x)$ is the gradient vector of the null density $f(x) = h(r)$. Thus, $f_0^{(1)}(x) = x\frac{h'(r)}{r}$.

We now make use of a general multidimensional integral formula in order to evaluate $\int |Z'x| f_0^{(1)}(x)dx$. The formula we use is the following:

For any two functions p, q such that the integrals below exist,

$$\int_{\mathcal{R}^p} p(z'x)q(||x||)dx = A_{p-1}(B)\int_0^\infty \int_0^\pi p(r||z||\cos\eta)(\sin\eta)^{p-2}r^{p-1}q(r)d\eta dr,$$

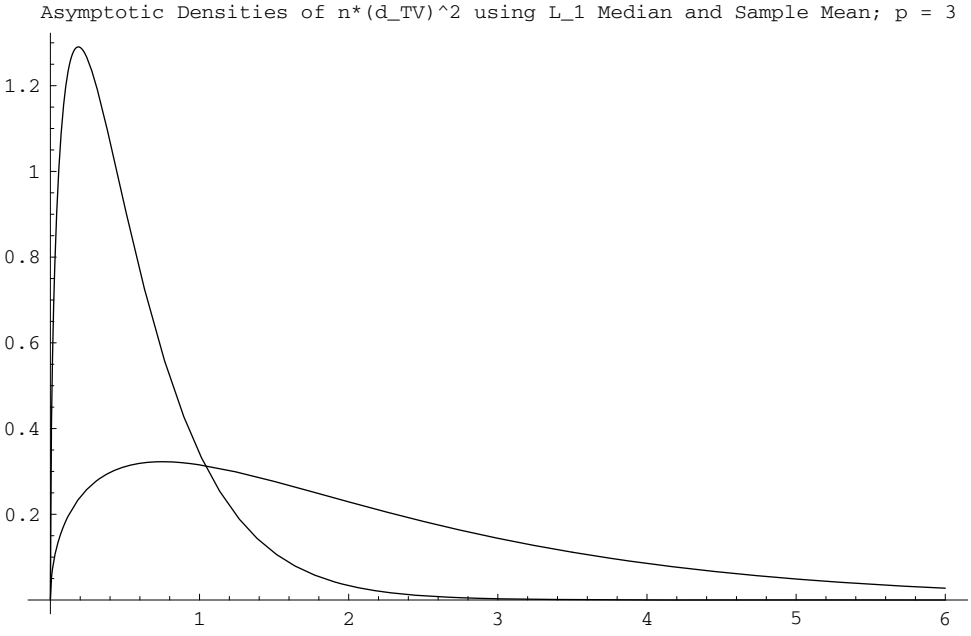
where $A_{p-1}(B) = \frac{2\pi^{\frac{p-1}{2}}}{\Gamma(\frac{p-1}{2})}$ is the surface area of the unit sphere B in $p-1$ dimensions. This formula results from a polar transformation of the rectangular coordinates $x_1, x_2, \ldots, x_p$.

Applying the formula to the function $p(z'x) = |z'x|$, and $q(r) = \frac{h'(r)}{r}$, we get

$$\int |Z' f_0^{(1)}(x)|dx = A_{p-1}(B)\int_0^\infty r^{p-1}|h'(r)|dr\left(\int_0^\pi |\cos\eta|(\sin\eta)^{p-2}d\eta\right)||Z||.$$

On the other hand, for the case of the L_1 median, $||Z||^2 \sim \frac{p}{(p-1)^2(E_F\frac{1}{||X||})^2}\chi^2(p)$, of which $E_F\frac{1}{||X||} = \frac{2\pi^{\frac{p}{2}}}{\Gamma(\frac{p}{2})}\int_0^\infty r^{p-2}h(r)dr$, by just a polar transformation. Plugging the respective expressions for $\int |Z' f_0^{(1)}(x)|dx$ and $\frac{p}{(p-1)^2(E_F\frac{1}{||X||})^2}$, on some algebra one gets the general result that for iid observations $X_1, X_2, \ldots, X_n$ from a spherically symmetric density $f(x-\theta) = h(||x-\theta||)$,

$$nd_{TV}^2(P_\theta, P_{\mathbf{M_n}}) \stackrel{\mathcal{L}}{\Rightarrow} \frac{p}{\pi(p-1)^4}\left(\frac{\Gamma(\frac{p}{2})}{\Gamma(\frac{p-1}{2})}\right)^2\left(\frac{\int_0^\infty r^{p-1}|h'(r)|dr}{\int_0^\infty r^{p-2}h(r)dr}\right)^2\chi^2(p).$$



Suppose now we make the additional assumption at this final stage that f is unimodal, so that $h(r)$ is decreasing and that f is bounded so that $h(0) < \infty$. Then, the final simplification comes from integrating $r^{p-1}|h'(r)|$ by parts, giving $\int_0^\infty r^{p-1}|h'(r)|dr = -\int_0^\infty r^{p-1}h'(r)dr = (p-1)\int_0^\infty r^{p-2}h(r)dr$, as the constant term vanishes due to the assumption $h(0) < \infty$. This simplification results in the fact that for all spherically symmetric unimodal densities in $\mathcal{R}^p$ with a bounded density, $nd^2_{TV}(P_\theta, P_{\mathbf{M_n}})$ has the same limiting distribution given by $\frac{p}{4\pi}\big(\frac{\Gamma(\frac{p}{2})}{\Gamma(\frac{p+1}{2})}\big)^2\chi^2(p)$. This is a very interesting robustness property of the L_1 median.

By an analogous calculation, for general spherically symmetric densities (without requiring unimodality),

$$nd^2_{TV}(P_\theta, P_{\overline{\mathbf{X}}}) \stackrel{\mathcal{L}}{\Rightarrow} \frac{2\pi^{\frac{3}{2}p-1}}{\Gamma(\frac{p}{2})(\Gamma(\frac{p+1}{2}))^2}\bigg(\int_0^\infty r^{p-1}|h'(r)|dr\bigg)^2\bigg(\int_0^\infty r^{p+1}h(r)dr\bigg)\chi^2(p).$$

Notice that, unlike the L_1 median, even if we assume unimodality, the limit distribution is not free of h.

The densities of the two limiting distributions corresponding to $\mathbf{M_n}$, and $\overline{\mathbf{X}}$ are plotted above for the underlying density $h(r) = ce^{-r}$, with c denoting the normalizing constant. We can see from the plot that $\mathbf{M_n}$ results in a stochastically smaller limit distribution. The advantage in using the L_1 median increases very significantly as the number of dimensions increases, but even for a small $p = 3$, the plot illustrates the huge advantage in using the L_1 median compared to the sample mean for this case. Of course, the choice of $h(r) = ce^{-r}$ was an artifact, and any other heavy-tailed choice would illustrate the result as well.

4. Increasing dimensions

Principally due to certain problems in applications to do with astronomy, network data, and genetics, there has been a tremendous increase in the interest in inference problems in very high dimensions in the last decade or so. Typically, in the mathematical formulation of these problems, the affine dimension p of the parameter space is allowed to depend on n, the sample size, and additionally, $p = p_n$ is assumed to converge to ∞ as $n \to \infty$. Nonparametrics in such very high dimensional problems continue to remain problematic, and there has been a reemergence of Gaussian parametrics in the context of these modern very high dimensional problems. There are other reasons that the increasing dimension or the infinite dimension Gaussian mean setup is important. Results in Ibragimov and Has'minskii (1977), Brown and Low (1996), Nussbaum (1996), Ingster (2001; 2002), among others, show that the infinite or the increasing dimension Gaussian mean problem relates in a fundamental way to various other nonparametric problems, such as nonparametric density estimation and nonparametric regression. A massive literature has already accumulated, and for the most part, the theoretical developments have focused on point or confidence estimation or hypothesis testing about the parameter vector. The literature is too huge to cite. A few key references are Donoho and Johnstone (1998), Johnstone (2003), Bickel and Li (2006), Bickel and Levina (2008), Bickel, Ritov, and Tsybakov (2009), Cai and Low (2005), Cai, Xu, and Zhang (2009), Hall and Jin (2008), Fan, Hall, and Yao (2007), Fan, Samworth, and Wu (2009), Wasserman (2006), and Wasserman and Roeder (2009). A particularly relevant recent reference is Johnstone (2010), which addresses various Bayesian asymptotics problems in the increasing dimension Gaussian setup.

We treat estimation of the entire density itself under such high or ultra high dimensions. By considering the L_1 risk in the density estimation, we can make statements simultaneously about estimating the probability of arbitrary Borel sets. The results will clearly bring out the advantages of regularization, the necessity for sparsity when the dimensions are ultra high, and the failure of ordinary maximum likelihood to provide even consistent density estimation when the dimensions cross a threshold rate of growth. The exact threshold will be explicitly pinned down; so will be the precise extent of sparsity needed in order that regularization can succeed when maximum likelihood fails. First order asymptotic theory will be established without leaving any open cases. We believe that this is the first formal development of theory for density estimation in high and ultra high dimensions.

Here is the setup and the notation that we will follow throughout this section. We have $X_1, X_2, \cdots, X_n$ iid $p = p_n$-dimensional Gaussian random vectors, each with mean vector θ and a known nonsingular covariance matrix Σ: $X_i \overset{iid}{\sim} N_p(\theta, \Sigma)$. The dependence of p (and θ and Σ) on n will be suppressed for notational ease. The common density function of our sample observations is $f(x|\theta) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\theta)'\Sigma^{-1}(x-\theta)}, x \in \mathcal{R}^p$. Our problem, as in the previous sections, is to estimate this density function itself, by using the sample data $X_1, X_2, \cdots, X_n$. Also, as in the previous sections, we continue to use the L_1 risk as our criterion. Because the covariance matrix Σ is assumed to be known, it may in fact be taken to be the just the $p \times p$ identity matrix, as can be quickly seen by making a linear transformation of the sample observations. This is a useful notational reduction, and Σ will be taken to be $I_{p\times p}$ in the rest of the section.

Here is some more notation that will be used for the rest of this section. ρ_n will denote the L_1 distance between $f(x|\theta)$ and $f(x|\hat{\theta})$ for a generic estimate sequence $\hat{\theta} = \hat{\theta}_n$; that is, $\rho_n = \int_{\mathcal{R}^p} |f(x|\theta) - f(x|\hat{\theta})|dx$. The MLE $\overline{X}$ will be denoted as $\hat{\theta}_{1n}$, and for regularization we use the canonical James-Stein estimator (James and Stein (1961)) $\hat{\theta}_{2n} = (1 - \frac{\alpha_n}{n\overline{X}'\overline{X}})\overline{X}$. The original choice of James and Stein was $\alpha_n = (p-2)$; we use $\alpha_n = p$, which leads to the same asymptotic results as for $\alpha_n = (p-2)$, but reduces the algebraic complexity of the calculations. Of course, regularization can be done by using various other estimates, such as the positive-part James-Stein estimator (see Strawderman (2000)), or hard thresholding regularization procedures, as in Donoho and Johnstone (1995). However, we consider only the canonical James-Stein estimator in this article due to reasons of space. Let

$$\sqrt{n}(\overline{X} - \theta) = Z_n; \qquad \frac{Z_n'Z_n - p}{\sqrt{2p}} = W_n. \tag{4.1}$$

Additionally, the notation W will be used to denote a generic standard normal variable. Note that for each fixed n, $Z_n'Z_n \sim \chi^2_p$; hence, as $n \to \infty, W_n \overset{\mathcal{L}}{\Rightarrow} W \sim N(0,1)$. In particular, W_n is an $O_p(1)$ sequence.

It is helpful to have a preview of what the entire set of results in this section says, which we present first before giving the theorems.

4.1. Preview of the results

(a) When $p \to \infty$, but at a slower rate than n, i.e., $p = o(n)$, both the MLE $\hat{\theta}_{1n}$ and the regularized estimate $\hat{\theta}_{2n}$ lead to consistent density estimation. Furthermore, the first order asymptotic theory for ρ_n, the L_1 risk, coincides (under a condition), and ρ_n goes to zero at the best possible rate, namely, $n^{-1/2}$. The interpretation is that in the *slowly increasing dimensional case*, regularization is not really necessary, and maximum likelihood itself does the job.

(b) If $p \sim n$, then maximum likelihood starts to falter and ρ_n (with $\hat{\theta}_n$ = MLE) does not even converge in probability to zero. We already lose consistency. If, in particular, $p = c^2 n$, then $\rho_n \overset{P}{\Rightarrow} 4[\Phi(\frac{c}{2}) - \frac{1}{2}]$, and $\sqrt{n}(\rho_n - 4[\Phi(\frac{c}{2}) - \frac{1}{2}]) \overset{\mathcal{L}}{\Rightarrow} N(0, 2\phi^2(\frac{c}{2}))$.

(c) If p grows faster than n, i.e., $\frac{p}{n} \to \infty$, then maximum likelihood completely falls apart and $\rho_n \overset{P}{\Rightarrow} 2$, *whatever be the parameter vector* θ. *Sparsity of the parameter vector is not going to save maximum likelihood in these ultra high dimensional cases.* Thus, $p \sim n$ is the threshold where maximum likelihood breaks down, and in the ultra high dimensional case, the true density and the density estimated by using maximum likelihood will sit essentially on disjoint subsets of $\mathcal{R}^p$.

(d) However, these problematic cases for the maximum likelihood estimate do not create problems for the regularized estimate, *as long as the parameter vector* θ *is just sufficiently sparse* (in the sense that $\|\theta\| \to 0$ at a suitable rate). Thus, regularization and sparsity join hands together to rescue the situation that maximum likelihood cannot deal with.

(e) Precisely, here is how the rate of growth of p and sparsity of the parameter vector together determine the asymptotic fate of the regularized density estimate:

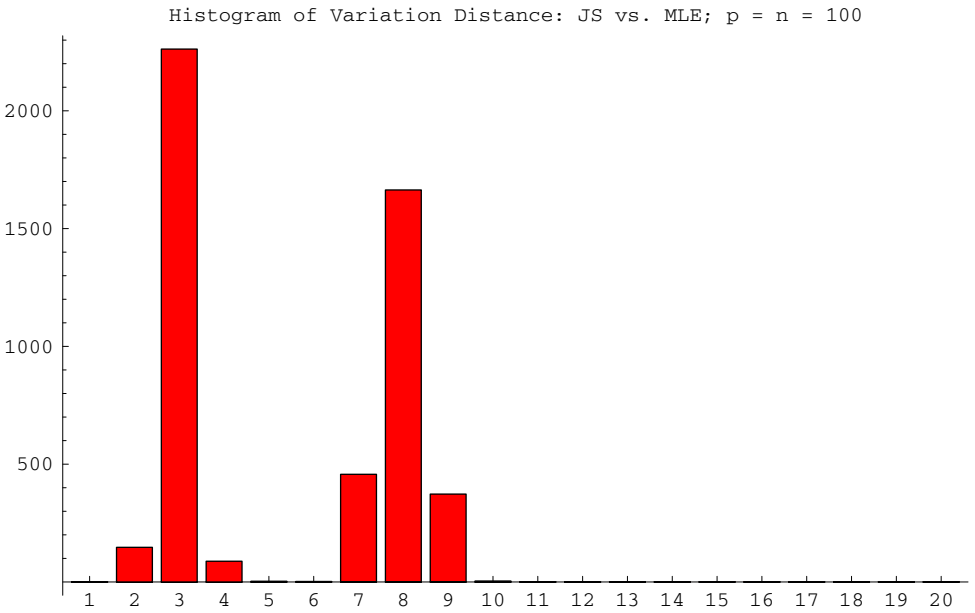
(i) If $\frac{n}{p}\theta'\theta \to 0$ and $n\theta'\theta \to \infty$, then $\frac{\rho_n}{||\theta||} \stackrel{P}{\Rightarrow} \sqrt{\frac{2}{\pi}}$.

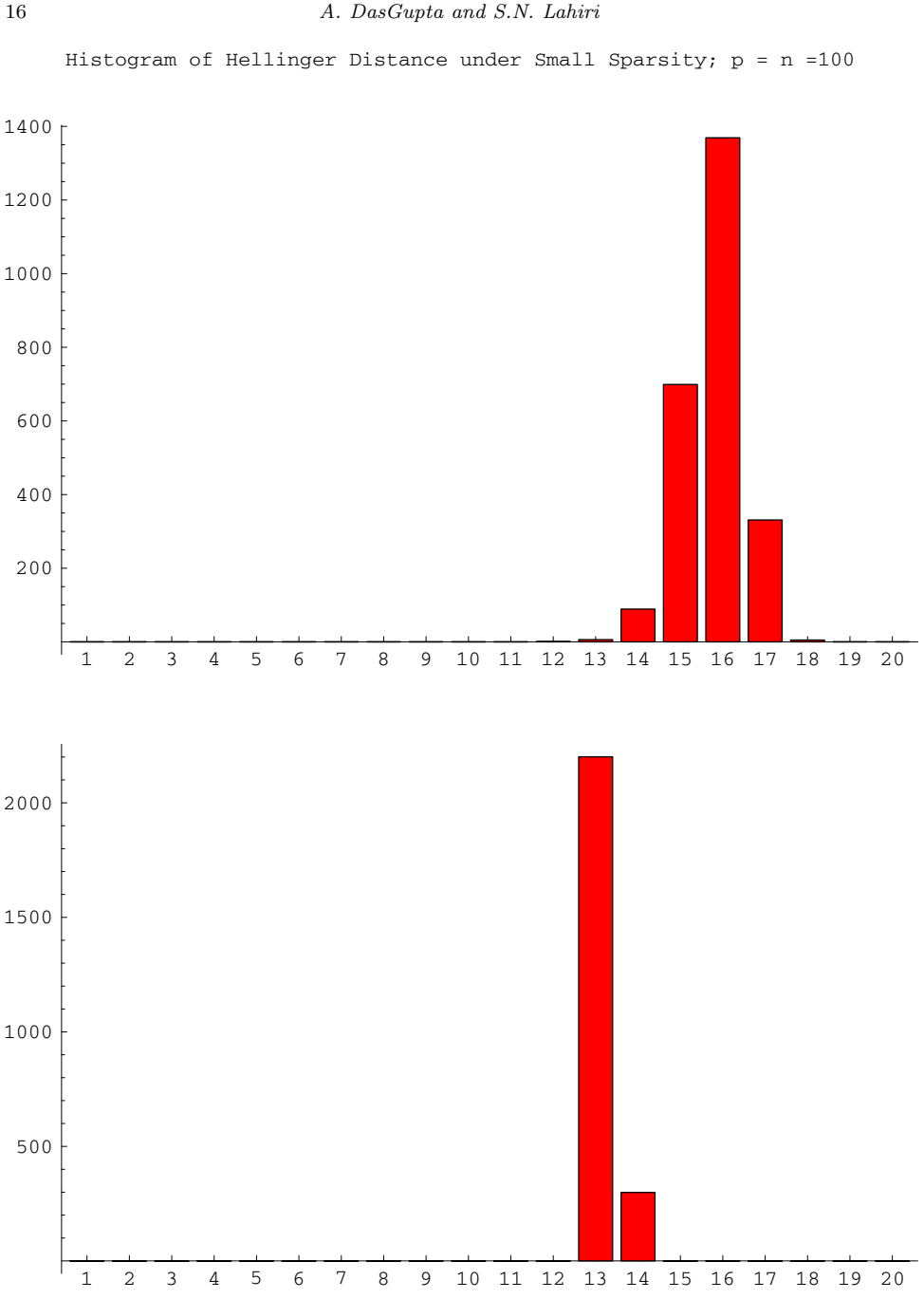
(ii) If $\frac{n}{p}\theta'\theta \to 0$ and $n\theta'\theta \to b^2$ $(0 < b < \infty)$, then $\sqrt{n}\rho_n \stackrel{\mathcal{L}}{\Rightarrow} \sqrt{\frac{2}{\pi}[b^2 + 2W^2]}$.

(iii) If $\frac{n}{p}\theta'\theta \to 0$ and $n\theta'\theta \to 0$, then $\sqrt{n}\rho_n \stackrel{\mathcal{L}}{\Rightarrow} \frac{2}{\sqrt{\pi}}|W|$.

Case (i) in (e) corresponds to the case where the dimensions are ultra high, and the parameter vector is not adequately sparse. In this case, ρ_n goes to zero at the slower rate of $||\theta||$, which is the loss in quality of estimation due solely to the lack of adequate sparsity. *Note the important fact that if* $||\theta|| \not\to 0$*, then in spite of regularization, we will lose consistency.* Cases (ii) and (iii) in (e) correspond to the adequately sparse case, and here we notice that the sparser the parameter vector is, i.e., the smaller b is, the smaller is the weak limit of ρ_n stochastically. *In addition, when we have this sort of adequate sparsity, we can still have convergence at the best possible* $n^{-1/2}$ *rate.* This is the exact gain in regularizing the estimate when we reach the ultra high dimensions.

As graphical illustration of the advantages of regularization, we provide below two histograms of simulated values of the variation distance ρ_n. In each plot, $p = n = 100$. In the first plot, $\theta_1 = \cdots = \theta_{10} = .1$, and the rest of the θ_i are zero. Thus, $\frac{n}{p}\theta'\theta = .1$ and $n\theta'\theta = 10$. This corresponds to medium sparsity. From the histogram, we see that the advantages of regularization are very substantial. If we do not regularize, ρ_n ranges between .33 and .48 (that is what the scale means), while if we do regularize, then ρ_n ranges between .08 and .23. The ranges are nonoverlapping. In the second plot, $\theta_1 = \cdots = \theta_{50} = .1$, and the rest of the θ_i are zero. Thus, $\frac{n}{p}\theta'\theta = .5$ and $n\theta'\theta = 50$. This corresponds to mild sparsity. From the histogram, we see that the advantages of regularization are still substantial, but now the ranges overlap. It appears that regularization would be a good idea when p is large even under mild sparsity.





4.2. The theorems

The results indicated in the preview above are proved in this section. The technique is to use an all at one time exact explicit formula for the L_1 risk ρ_n and then use it to carefully write stochastic expansions for ρ_n. Under the various configurations of $\frac{p}{n}$ and $\theta'\theta$, different terms in the stochastic expansion become the stochastically dominant term, which then determine the precise asymptotics. The exact formula for ρ_n is given first. The key to writing this formula is the convenient fact that

a linear function of a Gaussian vector is a univariate normal, which allows one to replace a seemingly complicated p-dimensional integral by a one dimensional integral, and it then turns out that this one dimensional integral can be done in closed form. Here is the formula.

Lemma 1. *Let $f(x|\theta)$ be the density of the $N_p(\theta, I_{p\times p})$ distribution. Then, for any estimate $\hat{\theta}$, the L_1 distance ρ_n admits the formula*

$$\rho_n = \int_{\mathcal{R}^p} |f(x|\theta) - f(x|\hat{\theta})|dx = 2\left[2\Phi(\frac{||\hat{\theta}-\theta||}{2}) - 1\right]. \tag{4.2}$$

Proof. By its definition,

$$\rho_n = \frac{1}{(2\pi)^{p/2}} \int_{\mathcal{R}^p} |e^{-\frac{1}{2}(x-\theta)'(x-\theta)} - e^{-\frac{1}{2}(x-\hat{\theta})'(x-\hat{\theta})}|dx$$

$$= \frac{1}{(2\pi)^{p/2}} \int_{\mathcal{R}^p} |e^{-\frac{1}{2}(y-(\hat{\theta}-\theta))'(y-(\hat{\theta}-\theta))} - e^{-\frac{1}{2}y'y}|dy$$

$$= \frac{e^{-\frac{1}{2}(\hat{\theta}-\theta)'(\hat{\theta}-\theta)}}{(2\pi)^{p/2}} \int_{\mathcal{R}^p} e^{-\frac{1}{2}y'y}|e^{(\hat{\theta}-\theta)'y} - e^{\frac{1}{2}(\hat{\theta}-\theta)'(\hat{\theta}-\theta)}|dy$$

$$= \frac{e^{-\frac{1}{2}(\hat{\theta}-\theta)'(\hat{\theta}-\theta)}}{\sqrt{2\pi}||\hat{\theta}-\theta||} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\frac{y^2}{||\hat{\theta}-\theta||^2}} |e^y - e^{\frac{||\hat{\theta}-\theta||^2}{2}}|dy$$

(this is the reduction to a one dimensional integral that was mentioned above)

$$= \frac{e^{-\frac{1}{2}(\hat{\theta}-\theta)'(\hat{\theta}-\theta)}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} |e^{||\hat{\theta}-\theta||z} - e^{\frac{||\hat{\theta}-\theta||^2}{2}}|dz$$

$$= e^{-\frac{1}{2}||\hat{\theta}-\theta||^2}\left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{||\hat{\theta}-\theta||}{2}} e^{-\frac{z^2}{2}}(e^{\frac{||\hat{\theta}-\theta||^2}{2}} - e^{||\hat{\theta}-\theta||z})dz\right.$$

$$\left. + \frac{1}{\sqrt{2\pi}} \int_{\frac{||\hat{\theta}-\theta||}{2}}^{\infty} e^{-\frac{z^2}{2}}(e^{||\hat{\theta}-\theta||z} - e^{\frac{||\hat{\theta}-\theta||^2}{2}})dz\right].$$

Each of these two integrals in the line above can be calculated in closed form, by completing the squares in the exponents in order to turn them into some other univariate normal density, and then write the integrals in terms of the standard normal CDF. Indeed, the two integrals are equal, and on simplification the $e^{-\frac{1}{2}||\hat{\theta}-\theta||^2}$ term outside the integrals cancels, resulting finally in the formula given in the statement of the lemma.

It is clear from this lemma that $\rho_n \stackrel{P}{\Rightarrow} 0$ iff $||\hat{\theta} - \theta|| \stackrel{P}{\Rightarrow} 0$. Since convergence in probability of a nonnegative sequence $\{Y_n\}$ is metrized by $E\frac{Y_n}{1+Y_n}$, we have the following interesting connection between estimation of f_θ and estimation of θ.

$$\rho_n \stackrel{P}{\Rightarrow} 0 \,\text{iff}\, E_\theta \frac{||\hat{\theta}-\theta||}{1+||\hat{\theta}-\theta||} \to 0.$$

In other words, the density estimation problem is equivalent to the estimation of the mean with the bounded loss function $\frac{||a-\theta||}{1+||a-\theta||}$. □

We now present the main theorems of this section.

Theorem 4.1 (Performance of Maximum Likelihood).

(a) (Slow growth of p). Suppose $p = o(n)$. Then under any sequence of mean vectors θ,

$$\sqrt{\frac{n}{p}}\rho_n \stackrel{P}{\Rightarrow} \sqrt{\frac{2}{\pi}}.$$

(b) (i) If $p = o(n^{2/3})$, then

$$\sqrt{p}\,[\sqrt{\frac{n}{p}}\rho_n - \sqrt{\frac{2}{\pi}}] \stackrel{\mathcal{L}}{\Rightarrow} N(0, \frac{1}{\pi});$$

(ii) If $\frac{p}{n^{2/3}} \to 1$, then

$$\sqrt{p}\,[\sqrt{\frac{n}{p}}\rho_n - \sqrt{\frac{2}{\pi}}] \stackrel{\mathcal{L}}{\Rightarrow} N(-\frac{1}{12\sqrt{2\pi}}, \frac{1}{\pi});$$

(iii) If $\frac{p}{n^{2/3}} \to \infty$, then

$$\sqrt{p}\,[\sqrt{\frac{n}{p}}\rho_n - \sqrt{\frac{2}{\pi}} + \frac{p}{n}\frac{1}{12\sqrt{2\pi}}] \stackrel{\mathcal{L}}{\Rightarrow} N(0, \frac{1}{\pi}).$$

(c) If $\frac{p}{n} \not\to 0$, then ρ_n does not converge in probability to zero.

(d) (The Boundary). Suppose $p = c^2 n$ for some fixed c. Then

$$\sqrt{n}\Big(\rho_n - 4[\Phi(\frac{c}{2}) - \frac{1}{2}]\Big) \stackrel{\mathcal{L}}{\Rightarrow} \sqrt{2}\phi(\frac{c}{2})W,$$

where $W \sim N(0,1)$.

(e) (Ultra High p). Suppose $\frac{p}{n} \to \infty$. Then

$$\rho_n \stackrel{P}{\Rightarrow} 2.$$

Proof. We use the formula of Lemma 1. When $\hat{\theta} = \overline{X}$, the MLE, using the notation defined at the beginning of this section,

$$||\hat{\theta} - \theta||^2 = \frac{Z_n' Z_n}{n} = \frac{p}{n}(1 + \sqrt{\frac{2}{p}}W_n)$$

$$\Rightarrow \rho_n = 4\Big[\Phi\Big(\frac{\sqrt{p}}{2\sqrt{n}}\sqrt{1 + \sqrt{\frac{2}{p}}W_n}\Big) - \frac{1}{2}\Big]$$

$$= 4\Big[\Phi(\frac{\sqrt{p}}{2\sqrt{n}} + \frac{W_n}{2\sqrt{2n}}) - \frac{1}{2}\Big] + O_p(\frac{1}{\sqrt{np}}). \tag{4.3}$$

This representation will be useful to us for establishing each part of this theorem. First, for part (a), from this last representation, when $p \to \infty, \frac{p}{n} \to 0$,

$$\rho_n = 4\Big[\Phi(0) + (\frac{\sqrt{p}}{2\sqrt{n}} + \frac{W_n}{2\sqrt{2n}})\phi(0) - \frac{1}{2}\Big] + O_p((\frac{p}{n})^{3/2}) + O_p(\frac{1}{\sqrt{np}})$$

$$\Rightarrow \frac{\sqrt{n}}{\sqrt{p}}\rho_n = 2\phi(0) + \frac{\sqrt{2}}{\sqrt{p}}\phi(0)W_n + O_p(\frac{p}{n}) + O_p(\frac{1}{p}) = 2\phi(0) + o_p(1).$$

Therefore, $\frac{\sqrt{n}}{\sqrt{p}}\rho_n \stackrel{P}{\Rightarrow} 2\phi(0) = \sqrt{\frac{2}{\pi}}$, which proves part (a).

By expanding $\sqrt{1+\sqrt{\frac{2}{p}}W_n}$ and also expanding $\Phi(z)$ around $z=0$, we get the stochastic expansion

$$\frac{\sqrt{n}}{\sqrt{p}}\rho_n - 2\phi(0) = \phi(0)\sqrt{\frac{2}{p}}W_n + \frac{\phi''(0)}{12}\frac{p}{n} + O_p(\frac{\sqrt{p}}{n})$$

$$\Rightarrow \frac{\sqrt{n}}{\sqrt{p}}\rho_n - 2\phi(0) = \phi(0)\sqrt{\frac{2}{p}}W_n + \frac{\phi''(0)}{12}\frac{p}{n} + o_p(\min(\frac{1}{\sqrt{p}}, \frac{\sqrt{p}}{n})).$$

Each result in part (b) follows from this last line, by considering the cases (i), (ii), (iii) separately. The detail is omitted due to limited space.

Part (c) also follows from the representation for ρ_n derived above. For, if $\frac{p}{n} \not\to 0$, there must be a subsequence $\frac{p_{n_j}}{n_j} \to c^2$ for some $c \in (0, \infty]$. The representation

$$\rho_n = 4\left[\Phi(\frac{\sqrt{p}}{2\sqrt{n}} + \frac{W_n}{2\sqrt{2n}}) - \frac{1}{2}\right] + O_p(\frac{1}{\sqrt{np}})$$

shows that along this subsequence, ρ_n converges in probability to $4\left[\Phi(\frac{c}{2}) - \frac{1}{2}\right]$, which is strictly positive for any $c \in (0, \infty]$.

The proof of part (d) is essentially the same as that of part (c). Writing $\frac{p}{n} = c^2$, the representation for ρ_n works out to

$$\rho_n = 4\left[\Phi(\frac{c}{2}) - \frac{1}{2} + \frac{W_n}{2\sqrt{2n}}\phi(\frac{c}{2})\right] + O_p(\frac{1}{n})$$

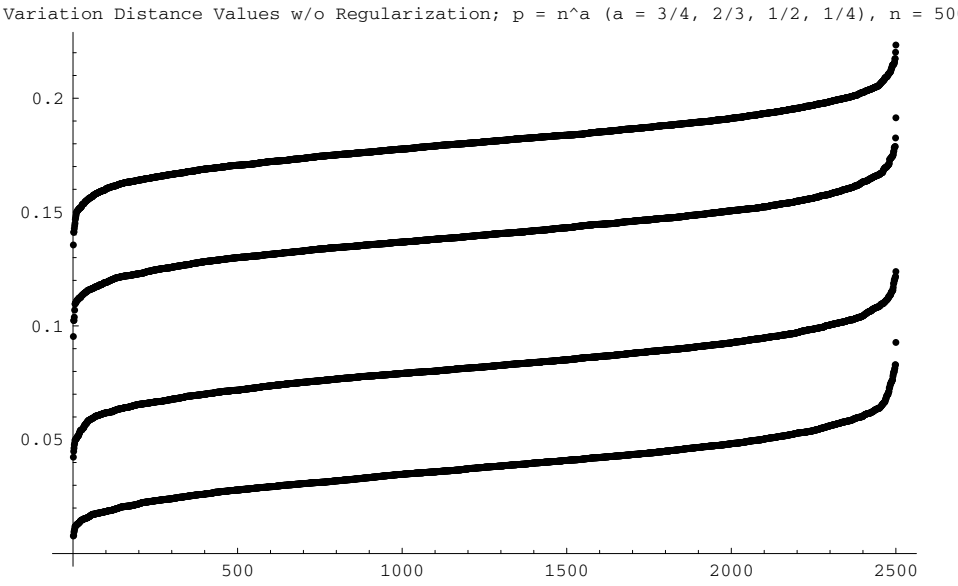
$$\Rightarrow \rho_n - 4\left[\Phi(\frac{c}{2}) - \frac{1}{2}\right] = \sqrt{\frac{2}{n}}\phi(\frac{c}{2})W_n + O_p(\frac{1}{n}).$$

Since $W_n \stackrel{\mathcal{L}}{\Rightarrow} W \sim N(0,1)$, the result of part (d) follows.

Part (e) follows immediately from the same representation for ρ_n. This completes the proof of this theorem. □

Remarks. It should be pointed out that if in part (d) we only assume that $\frac{p}{n} \to c^2 (0 < c < \infty)$, then the final result will depend on the exact rate of convergence of $\frac{p}{n} - c^2$ to zero. That is the reason that $\frac{p}{n}$ was taken as equal to c^2 in part (d). An important special case is $p = n$, and part (d) tells us that if $p = n$, $\sqrt{n}(\rho_n - .7659) \stackrel{\mathcal{L}}{\Rightarrow} N(0, .248)$.

Part (a) of Theorem 4.1 says that if $p = o(n)$, then $\rho_n \stackrel{P}{\Rightarrow} 0$. Thus, theoretically, if p grows slower than n, then for large n, ρ_n should be small and ordinary maximum likelihood ought to work fine. This can be a bit misleading in practice. The following plot gives the plots of 2500 simulated values of $\frac{1}{2}\rho_n$ when $n = 500$ and p is, from top to bottom, $n^{3/4}, n^{2/3}, \sqrt{n}$, and $n^{1/4}$. For example, when $p = n^{3/4} = 105$ (the top curve), we see that the values of $\frac{1}{2}\rho_n$ vary between .13 and .23. This is unsatisfactory. Only for the lowest curve ($p = n^{1/4}$), the values are at all acceptably small. Thus, there can be a divide between practical performance and predicted performance, as asserted in the theorem, unless n is very very large. The theorem gives qualitative insight, but it should not be taken literally.



The next result derives the asymptotics of the regularized estimate. *The various parts of this next theorem show how regularization can help in achieving consistency in high or even ultra high dimensions, as long as the ill effects of the high dimensionality are neutralized by the right amount of sparsity in a delicate balancing act. The theorem reinforces the inevitable need for regularization when the curse of dimensionality strikes.*

Theorem 4.2 (Performance of the Regularized Estimate).

(a) (Slow growth of p). Suppose $p = o(n)$ *and that* $\frac{n}{p}(\theta'\theta) \to \infty$. *Then*

$$\sqrt{\frac{n}{p}}\rho_n \stackrel{P}{\Rightarrow} \sqrt{\frac{2}{\pi}}.$$

(b) (Slow growth of p plus some sparsity). Suppose $p = o(n)$ *and that* $\frac{n}{p}(\theta'\theta) \to a^2$ *for some* $a \in (0, \infty)$. *Then*

$$\sqrt{\frac{n}{p}}\rho_n \stackrel{P}{\Rightarrow} \frac{a\sqrt{2}}{\sqrt{\pi(1+a^2)}}.$$

(c) (Connection to Pinsker's Theorem). Suppose $p = n$ *and* $\theta'\theta \to c^2 (0 < c < \infty)$. *Then*

$$\rho_n \stackrel{P}{\Rightarrow} 2\left[2\Phi(\frac{c}{2\sqrt{1+c^2}}) - 1\right].$$

(d) (Arbitrary growth of p).

(i) (Some sparsity). Suppose $\frac{n}{p}(\theta'\theta) \to 0$, *but* $n(\theta'\theta) \to \infty$. *Then*

$$\frac{1}{||\theta||}\rho_n \stackrel{P}{\Rightarrow} \sqrt{\frac{2}{\pi}}.$$

(ii) (Medium sparsity). Suppose $\frac{n}{p}(\theta'\theta) \to 0$, *and* $n(\theta'\theta) \to b^2$ *for some* $b \in (0, \infty)$. *Then*

$$\sqrt{n}\rho_n \stackrel{\mathcal{L}}{\Rightarrow} \sqrt{\frac{2}{\pi}(b^2 + 2W^2)},$$

where $W \sim N(0,1)$.

(iii) (High sparsity). Suppose $n(\theta'\theta) \to 0$ *(and hence* $\frac{n}{p}(\theta'\theta)$ *automatically goes to zero). Then*

$$\sqrt{n}\rho_n \stackrel{\mathcal{L}}{\Rightarrow} \frac{2}{\sqrt{\pi}}|W|,$$

where $W \sim N(0,1)$.

Proof. Once again, we use the notation outlined at the beginning of this section and we use the explicit formula for ρ_n derived in Lemma 1. The estimate of θ in this theorem is the James-Stein estimate

$$\hat{\theta} = (1 - \frac{\alpha_n}{n\overline{X}'\overline{X}})\overline{X}.$$

This gives, on some algebra, which is omitted here,

$$||\hat{\theta} - \theta||^2 = ||\overline{X} - \theta||^2 - \frac{2\alpha_n \overline{X}'(\overline{X} - \theta)}{n\overline{X}'\overline{X}} + \frac{\alpha_n^2 \overline{X}'\overline{X}}{n^2(\overline{X}'\overline{X})^2}$$

$$= \frac{Z_n'Z_n}{n} + \frac{\alpha_n^2 - 2n\alpha_n(\theta + \frac{Z_n}{\sqrt{n}})'\frac{Z_n}{\sqrt{n}}}{n^2(\theta + \frac{Z_n}{\sqrt{n}})'(\theta + \frac{Z_n}{\sqrt{n}})}$$

This expression is valid for any choice of α_n. If we use $\alpha_n = p$, this becomes

$$||\hat{\theta} - \theta||^2 = \frac{p + \sqrt{2p}W_n}{n} - \frac{p^2 + 2p\sqrt{n}(\theta'Z_n) + 2\sqrt{2}p^{3/2}W_n}{n^2(\theta'\theta) + 2n^{3/2}(\theta'Z_n) + np + \sqrt{2p}nW_n}. \tag{4.4}$$

The task is now to identify the dominant terms in this stochastic representation under the various parts of this theorem, which then determine the asymptotics. We will provide here the derivations only for part (a) and part (d)(ii) due to space reasons.

First consider part (a). In this case, ultimately the term $\frac{p+\sqrt{2p}W_n}{n}$ will turn out to be the dominant term in $||\hat{\theta} - \theta||^2$, which is what happened in part (a) of Theorem 4.1, and that is why the result is also the same as in part (a) of Theorem 2.1. The reason that $\frac{p+\sqrt{2p}W_n}{n}$ is the dominant term is that the second term $\frac{p^2+2p\sqrt{n}(\theta'Z_n)+2\sqrt{2}p^{3/2}W_n}{n^2(\theta'\theta)+2n^{3/2}(\theta'Z_n)+np+\sqrt{2p}nW_n}$ is $o_P(\frac{p}{n})$. This requires some careful bookkeeping. We show that in this expression p^2 is the dominant term in the numerator and $n^2(\theta'\theta)$ is the dominant term in the denominator. For example, in the denominator,

$$\frac{np + \sqrt{2p}nW_n}{n^2(\theta'\theta)} = O_p(\frac{np}{n^2(\theta'\theta)}) = O_p(\frac{1}{\frac{n}{p}(\theta'\theta)}) = o_P(1).$$

Next, by the Cauchy-Schwarz inequality,

$$\frac{\left(n^{3/2}(\theta'Z_n)\right)^2}{n^4(\theta'\theta)^2} = O_P\left(\frac{n^3(Z_n'Z_n)}{n^4(\theta'\theta)}\right) = O_P\left(\frac{n^3 p}{n^4(\theta'\theta)}\right) = O_P\left(\frac{1}{\frac{n}{p}(\theta'\theta)}\right) = o_P(1).$$

On the other hand, in the numerator, $p^{3/2}W_n = o_P(p^2)$ because $W_n = O_P(1)$, and $p\sqrt{n}(\theta'Z_n) = o_P(p^2)$ by another application of the Cauchy-Schwarz inequality, Hence, putting these all together,

$$(4.5)\quad \frac{p^2+2p\sqrt{n}(\theta' Z_n)+2\sqrt{2}p^{3/2}W_n}{n^2(\theta'\theta)+2n^{3/2}(\theta' Z_n)+np+\sqrt{2p}nW_n} = O_P(\frac{p^2}{n^2(\theta'\theta)}) = O_P(\frac{p}{n}\frac{1}{\frac{n}{p}(\theta'\theta)}) = o_P(\frac{p}{n}),$$

because $\frac{n}{p}(\theta'\theta) \to \infty$ by assumption in part (a). This proves the result of part (a).

Consider next part (d)(ii) of the theorem. In this case, unlike part (a), the term $\frac{p+\sqrt{2p}W_n}{n}$ is no longer negligible, and we must deal with the entire expression for $||\hat{\theta}-\theta||^2$. Indeed, on a few lines of algebra and careful cancellations, the expression for $n||\hat{\theta}-\theta||^2$ reduces to

$$(4.6)\quad n||\hat{\theta}-\theta||^2 = \frac{2pW_n^2+np(\theta'\theta)+\sqrt{2p}(n\theta'\theta)W_n+2\sqrt{2}\sqrt{np}(\theta' Z_n)W_n}{(n\theta'\theta)+2\sqrt{n}(\theta' Z_n)+p+\sqrt{2p}W_n}.$$

It will now ultimately turn out that $2pW_n^2+np(\theta'\theta)$ is the dominant term in the numerator, and p is the dominant term in the denominator. Therefore, the asymptotics will be determined by those of $\frac{2pW_n^2+np(\theta'\theta)}{p} = 2W_n^2+b^2$, due to the assumption in part (c)(ii) that $n(\theta'\theta) \to b^2$. Once we have this, the result of part (d)(ii) falls out immediately by invoking our formula for ρ_n given in Lemma 1.

The verifications that $2pW_n^2+np(\theta'\theta)$ and p are the respective dominant terms in the numerator and the denominator are not difficult. The only term in the denominator that requires checking is $2\sqrt{n}(\theta' Z_n)$. Again, by the Cauchy-Schwarz inequality,

$$(\sqrt{n}(\theta' Z_n))^2 = O_P(n\theta'\theta(Z_n' Z_n)) = O(1)O_P(p) = O_P(p),$$

as was required. In the numerator, the term $\sqrt{2p}(n\theta'\theta)W_n$ is disposable because $n\theta'\theta = O(1)$ by assumption and $W_n = O_p(1)$. Likewise, the term $2\sqrt{2}\sqrt{np}(\theta' Z_n)W_n$ is also negligible by an application of the Cauchy-Schwarz inequality and the fact that $W_n = O_P(1)$. This completes our proof of part (d)(ii) of the theorem. □

References

Bickel, P. J. and Li, Bo. (2006). Regularization in statistics, With comments and a rejoinder by the authors, TEST, 15, 271–344.

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices, Ann. Statist., 36, 199–227.

Bickel, P. J., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of lasso and the Dantzig selector, Ann. Statist., 37, 1705–1732.

Brown, B. M. (1983). Statistical uses of the spatial median, J. Royal Stat. Society, Ser. B, 45, 1, 25–30.

Brown, L. and Low, M. (1996). Asymptotic equivalence of nonparametric regression and white noise, Ann. Statist., 24, 2384–2398.

Brown, L., Cai, T. and DasGupta, A. (2003). Interval estimation in Exponential families, Statist. Sinica, 13, 19–49.

Brown, M. (1995). The distribution of the total variation distance, with applications to simultaneous confidence intervals, Comput. Oper. Res., 22, 4, 373–381.

Cai, T. and Low, M. (2005). Nonparametric estimation over shrinking neighborhoods: Superefficiency and adaptation, Ann. Statist., 33, 184–213.

Cai, T., Xu, G. and Zhang, J. (2009). On recovery of sparse signals via ℓ_1 minimization, IEEE Trans. Inform. Theory, 55, 3388–3397.

Devroye, L. and Györfi, L. (1984). Nonparametric Density Estimation: The $L1$ View, Wiley, New York.

Diaconis, P. and Zabell, S. (1991). Closed form summation for classical distributions: variations on a theme of de Moivre, Statist. Sc., 6, 3, 284–302.

Donoho, D. and Johnstone, I. (1995). Adapting to unknown smoothness via wavelet shrinkage, Jour. Amer. Statist. Assoc., 90, 1200–1224.

Donoho, D. and Johnstone, I. (1998). Minimax estimation via wavelet shrinkage, Ann. Statist., 26, 879–921.

Fan, J. Q., Hall, P. and Yao, Q. (2007). To how many simultaneous hypothesis tests the normal, Student's t, or bootstrap calibration be applied?, Jour. Amer. Statist. Assoc., 102, 1282–1288.

Fan, J. Q., Samworth, R. and Wu, Y. (2009). Ultrahigh dimensional feature selection: Beyond the linear model, Jour. Machine Learning Res., 10, 2013–2038.

Hall, P. (1984). Central limit theorem for integrated squared error of multivariate nonparametric density estimators, Jour. Mult. Analysis, 14, 1, 1–16.

Hall, P. and Jin, J. (2008). Properties of higher criticism under strong dependence, Ann. Statist., 36, 381–402.

Hettmansperger, T. P. and McKean, J. W. (1998). *Robust Nonparametric Statistical Methods*, Kendall's Library of Statistics, John Wiley, New York.

Ibragimov, I. A. and Has'minskii, R. Z. (1977). On the estimation of an infinite-dimensional parameter in Gaussian white noise, Soviet Math. Doklady, 236, 1053–1055.

Ingster, Y. I. (2001). Adaptive estimation of a signal of growing dimension, I, Math. Methods of Statist., 10, 395–421.

Ingster, Y. I. (2002). Adaptive estimation of a signal of growing dimension, II, Math. Methods of Statist., 11, 37–68.

James, W. and Stein, C. (1961). Estimation with Quadratic Loss, Proc. Fourth Berkeley Symp. Math. Statist. Prob., I, 361–379.

Johnstone, I. (2003). Function Estimation in Gaussian Noise, Stanford University Mimmeographed Notes.

Johnstone, I. (2010). High dimensional Bernstein-von Mises: Simple examples, To appear, IMS Coll., Festschrift in Honor of Lawrence D. Brown, J. Berger, Tony Cai, I. Johnstone eds.

Le Cam, L. (1990). On the standard asymptotic confidence sets of Wald, Internat. Statist. Rev., 58, 2, 129–152.

Liu, H., Lafferty, J. and Wasserman, L. (2007). Sparse nonparametric density estimation in high dimensions using the *Rodeo*, Eleventh International Conference on Artificial Intelligence and Statistics, Vol. 2, 283–290.

Nussbaum, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise, Ann. Statist., 24, 2399–2430.

Resnick, S. I. (1987). *Extreme Values, Regular Variation, and Point Processes*, Springer-Verlag, New York.

Small, C. G. (1990). A survey of multidimensional medians, Internat. Statist. Rev., 58, 3, 263–277.

Strawderman, W. E. (2000). Minimaxity, Jour. Amer. Statist. Assoc., 95, 1364–1368.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large, Trans. Amer. Math. Soc., 54, 426–482.

Wasserman, L. (2006). All of Nonparametric Statistics, Springer, New York.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection, Ann. Statist., 37, 2178–2201.