SOME RESULTS ON THE CURSE OF DIMENSIONALITY AND SAMPLE SIZE RECOMMENDATIONS

by

Anirban DasGupta
Purdue University

Technical Report #97-11

Department of Statistics
Purdue University
West Lafayette, IN USA

July 1997

SOME RESULTS ON THE CURSE OF DIMENSIONALITY AND SAMPLE SIZE RECOMMENDATIONS

by

Anirban DasGupta*
Purdue University

Abstract

Multivariate density estimation is well known to be a tremendously difficult problem due to the occurrence of phenomena variously known as the corner effect and the curse of dimensionality. Specifically, histogram density estimation in high dimensions is plagued by the consequence that sampled observations tend to reside with high probability in low density regions of the sample space. In this article we attempt to quantify two central things: in how many dimensions, one starts to really feel the curse of dimensionality, and what sort of sample sizes are needed to do any kind of a reasonable inference in various dimensions. These questions cannot be formulated in a unique way. So the attempt is to derive a broad spectrum of results, which are then illustrated by extensive computation. A number of results may be of independent interest in combinatorics and applied probability. Our subjective conclusion after these extensive computations is that in 3 dimensions one often sees the most drastic effect relative to just one less dimension; in 5 dimensions one feels the curse of high dimensions rather strongly; in 10 dimensions, the feasibility of inference with realistic sample sizes basically vanishes. We also give a subjective minimum sample size recommendation based on the number of dimensions. These calculations are different in character from Epanechnikov(1969).

1. Introduction

Nonparametric density estimation has been a very active area of theoretical research for over forty years, starting with the striking result of Rosenblatt (1956) that a UMVUE of the density at a point does not exist. Silverman (1986), Devroye and Gyorfi (1984), Nadaraya (1983), Izenman (1992) and Scott (1992) describe various aspects and methods of density estimation, including the modern kernel estimation methods; one should also

* Research supported by National Security Agency Grant MDA 904–97–1–0031.

see the lucid review articles Wertz (1978) and Scott and Wand (1991).

Multivariate density estimation, despite an existing theory, is generally a very difficult exercise and plagued by a phenomenon commonly known as **the corner effect or curse of dimensionality**; see Scott (1992). The intrinsic problem is that sampled observations in high dimensions almost exclusively reside in the far corners or low density regions of the sample space and it is indeed difficult to catch any subtle features of the underlying density function except when one has the luxury of taking a huge number of observations. See Epanechnikov (1969) and Table 3.7 in Scott (1992). For some elegant probabilistic calculations on the phenomenon of corner effect, one may see Silverman (1986), Wegman (1990) and Section 1.5 in Scott (1992). The general goal of this article is to provide theoretical results on this phenomenon in greater generality (i.e., with less structural assumptions), to investigate the issues of how many dimensions does it take to feel the curse of dimensionality and to give some absolutely minimal sample size recommendations. There can be no unique answers to the questions we address and so the intention is to give theoretical calculations on various different formulations and try to see if there is a common thread in these calculations. The flavor is a bit more probabilistic than statistical, although statistical examples are given as well.

In Section 2, we demonstrate the phenomenon of corner effect in some generality by two results and an example. One result says that under a local Lipschitz condition on the least radial majorants of a sequence of densities $f_n(\underset{\sim}{x})$ in n-space, the probability of a fixed neighborhood goes to zero; this general result is then illustrated by two examples. A second result calculates the amount of a cube occupied by the inscribed L_p ball in n-space for a general $p > 0$, and examines the smallest dimension at which 90% or more of the cube is outside of the inscribed L_p ball. These are not new phenomena, but generalize the known cases quite well.

In Section 3, we build on the results of Section 2 by presenting an interesting question. The question is this: if k observations are taken **from the unit cube** in n-space, and we plant a "random set" in the cube, what is the probability that this set is entirely unvisited by the k observations? We take the set to be a random L_p ball, appropriately defined. For a general theory of random sets, one may consult Kendall (1974) and Matheron (1975), among many references. The mathematical formulae are then illustrated by computation and we again examine the smallest dimension at which the random set remains unvisited with a probability of 90% or more. We also examine the smallest sample size which will ensure that a random sphere gets visited with a probability of 50% or more. In this calculation, we find the astonishing fact that in 4 dimensions, we require 145,000 observations for this purpose! These calculations are new.

Having shown in Sections 2 and 3 that with a near certainty, sampled observations in a high dimensional space will not be visible in fixed neighborhoods, in Section 4 we try to further quantify the sample sizes that will be necessary so a number of fixed neighborhoods will be visited. Specifically, given observations $\underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots$ from a density $f(\underset{\sim}{x})$ in n-space, and a partition $\{S_1, S_2, \ldots, S_m\}$ of the sample space, we consider the minimum sample size N necessary to have each set $S_1, \ldots, S_m$ visited at least once. This therefore relates to the classic coupon collector's problem (see Holst (1986)). As illustration, we take the sets $S_1, S_2, \ldots, S_m$ to be the annuli $||\underset{\sim}{X}|| \leq 1, \ldots, m-2 < ||\underset{\sim}{X}|| \leq m-1, ||\underset{\sim}{X}|| > m-1$, and study N for the **multivariate normal and the multivariate t case**. These calculations are different in character from Epanechnikov (1969).

The article closes in Section 5 with some results on the interesting random variable $T = T(k, m, n) = \#$ elements of a partition $\{S_1, S_2, \ldots, S_m\}$ that remain unvisited after k sample observations $\underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots, \underset{\sim}{x}_k$ from a density of $f(\underset{\sim}{x})$ in n-space. We study $P(T > 0)$ and the distribution of T. When the elements of the partition have equal probability $\frac{1}{m}$, the distribution of T follows from consideration of Stirling's second numbers. In addition, we present some simulations on how T is affected if the members of the partition are not equally likely; the illustrative computing is for the multivariate normal case with the sets S_i being annuli.

In summary, the main results are therefore the following:

a. We give some general results on the presence of corner effect and the curse of dimensionality by making less structural assumptions. Here we can have the density $f_n(x)$ to depend on the dimension and even have the coordinates of $\underset{\sim}{X}$ collapse to zero at appropriate rates;

b. We give some results on the probability with which a random set placed in a cube stays unvisited;

c. We try to quantify the sample sizes necessary so fixed neighborhoods do get visited. This is illustrated by the case of the multivariate t distributions when the neighborhoods are the spherical annuli $||\underset{\sim}{X}|| \leq 1$, $||\underset{\sim}{X}|| \leq 2, \ldots, ||\underset{\sim}{X}|| > m$;

d. We also study the number of unvisited members of a partition of the sample space after a fixed number of observations;

e. In these results, we investigate how many dimensions it takes to feel the curse of dimensionality, by extensive computing;

f. We give some minimal sample size recommendations for various dimensions; these are subjective.

2. Two Illustrative Results

2.1. Corner Effect. It is well known that the amount of a cube occupied by a sphere in n-space goes to zero as $n \to \infty$. We first give a result that helps answer the following more general question: if B is a general inscribed L_p ball, $p > 0$, how many dimensions does it take for the fraction of the cube occupied by B to be less than a given ϵ?

Proposition 1. Let C be the unit n-dimensional cube $C = \{\underset{\sim}{x}\colon \max\limits_i |x_i| \leq 1\}$ and for $p > 0$, B_p the inscribed L_p ball $B_p = \{\underset{\sim}{x}\colon \sum\limits_i |x_i|^p \leq 1\}$. Let $a = a(p) = \left(\frac{e}{p}\right)^{\frac{1}{p}} \Gamma\left(\frac{1}{p}+1\right)$ and $b = b(p) = \frac{\varepsilon\sqrt{2\pi}}{\sqrt{p}}$ for any given $\varepsilon > 0$. Then $\varepsilon(n,p) = \frac{\text{vol }(B_p)}{\text{vol }(C)} < \varepsilon$ if $\log n > p\left(\log\ a - \frac{\log\ b}{n} - \frac{\log\ n}{2n}\right)$.

Proof: For any $p > 0$, the volume of the unit L_p ball in n dimensions equals

$$v(n,p) = \left(\frac{2}{p}\right)^n \frac{\left(\Gamma(\frac{1}{p})\right)^n}{\Gamma\left(\frac{n}{p}+1\right)} = \frac{2^n\left(\Gamma(\frac{1}{p}+1)\right)^n}{\Gamma\left(\frac{n}{p}+1\right)}, \tag{1}$$

and hence

$$\varepsilon(n,p) = \frac{\left(\Gamma(\frac{1}{p}+1)\right)^n}{\Gamma\left(\frac{n}{p}+1\right)}. \tag{2}$$

Using the fact that $\Gamma(x+1) \geq e^{-x}x^{x+\frac{1}{2}}\sqrt{2\pi}$, from (2) one has

$$\begin{aligned} \varepsilon(n,p) &\leq \left(\frac{e^{\frac{1}{p}}\Gamma(\frac{1}{p}+1)}{n^{\frac{1}{p}}p^{\frac{1}{p}}}\right)^n \sqrt{\frac{p}{2n\pi}} \\ &= \left(\frac{a}{n^{\frac{1}{p}}}\right)^n \frac{\varepsilon}{b\sqrt{n}} \\ &< \varepsilon \end{aligned} \tag{3}$$

if $\left(\frac{a}{n^{\frac{1}{p}}}\right)^n \frac{1}{b\sqrt{n}} < 1$, which simplifies to $\log\ n > p(\log\ a - \frac{\log\ b}{n} - \frac{\log\ n}{2n})$.

Remark. Table 1 lists, for $p = .5, 1, 2, 3$ and 4, the threshold value n at which the ratio $\varepsilon(n,p) \leq .1$.

Table 1

Minimum # dimensions at which 90%
of a cube is outside of an L_p ball

	p				
	.5	1	2	3	4
$\underline{n}$ (dimensions)	3	4	5	7	11

So perhaps one can say that the phenomenon of corner effect manifests itself strongly in 5 or more dimensions.

2.2 Curse of Dimensionality. The next result says that with very little structural assumptions on a sequence of n-dimensional densities f_n, sample observations will tend to reside in low density regions of the sample space. This generalizes previous calculations in Silverman (1986), Scott (1992) and Wegman (1990). We first give a definition.

Definition. Let $f(\underset{\sim}{x})$ be a given function on $\mathbb{R}^n$. The least radial majorant of f is the function $g(||\underset{\sim}{x}||) = \sup\limits_{\underset{\sim}{y}:\ ||\underset{\sim}{y}||=||\underset{\sim}{x}||} f(\underset{\sim}{y})$.

Theorem 1. Let $f_n(\underset{\sim}{x})$ be a sequence of densities in $\mathbb{R}^n$ and $g_n(||\underset{\sim}{x}||)$ the corresponding sequence of least radial majorants. Let $M_n = \sup\limits_{r>0} |g_n(r) - g_n(1)|/|r-1|$ (We define 0/0 to be 1). Suppose g_n satisfies the following two conditions:

$\underline{\text{i}}$ $\frac{\sqrt[n]{g_n(1)}}{\sqrt{n}} = o(1)$ (4)

ii $\frac{\sqrt[n]{M_n}}{\sqrt{n}} = o(1).$ (5)

Then for any fixed $k > 0$, $P_{f_n}(||\underset{\sim}{X}|| \leq k) \to 0$.

Before giving the proof of Theorem 1, we give the following corollary. It says that for a multivariate normal distribution, probabilities of fixed neighborhoods go to zero even if the covariance matrix is allowed to depend on the dimension and collapses to zero at an appropriate rate.

Corollary 1. Let $\underset{\sim}{X}_n \sim N_n(\underset{\sim}{0}, \Sigma_n)$. If $n\lambda_{\text{min}}(\Sigma_n) \to \infty$ and $\liminf \sqrt[n]{\lambda_{\text{max}}(\Sigma_n)} > 0$, then $P(||\underset{\sim}{X}_n|| \leq k) \to 0$ for any fixed k (here λ_{min} and λ_{max} denote the minimum and the maximum eigenvalues of Σ_n). In particular, if $\Sigma_n = \sigma_n^2 I_n$, then $P(||\underset{\sim}{X}_n|| \leq k) \to 0$ if $\sigma_n^2 \to 0$ at any rate slower than $\frac{1}{n}$, i.e., if $n\sigma_n^2 \to \infty$.

Proof of Corollary 1: Here $f_n(\underset{\sim}{x}) = {\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_n|^{\frac{1}{2}}}}^{-\frac{1}{2}\underset{\sim}{x}'\Sigma_n^{-1}\underset{\sim}{x}}$, and hence

$$g_n(r) = \frac{1}{(2\pi)^{n/2}|\Sigma_n|^{1/2}} e^{-\frac{1}{2\lambda_{\text{max}}(\Sigma_n)}r^2}. \tag{6}$$

In particular,

$$\begin{aligned} g_n(1) &\leq \frac{1}{(2\pi)^{n/2}\lambda_{\text{min}}^{n/2}(\Sigma_n)} \\ \Rightarrow \frac{\sqrt[n]{g_n(1)}}{\sqrt{n}} &\leq \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{n\lambda_{\text{min}}(\Sigma_n)}} \\ &\to 0 \text{ by hypothesis.} \end{aligned}$$

Secondly,

$$\begin{aligned} M_n &= \sup_{r\geq 0} \frac{|g_n(r) - g_n(1)|}{|r-1|} \\ &\leq \sup_{s\geq 0} |g_n'(s)| \\ &= \sup_{s\geq 0} \frac{1}{(2\pi)^{n/2}|\Sigma_n|^{1/2}} \cdot \frac{s}{\lambda_{\text{max}}(\Sigma_n)} \cdot e^{-\frac{1}{2\lambda_{\text{max}}(\Sigma_n)}s^2} \\ &\leq \frac{B}{(2\pi)^{n/2}|\Sigma_n|^{1/2}\sqrt{\lambda_{\text{max}}(\Sigma_n)}} \\ &\quad \text{(here } B \text{ is a universal constant, as the function} \\ &\quad ze^{-\frac{z^2}{2}} \text{ is uniformly bounded)} \\ &\leq \frac{B}{(2\pi)^{n/2}\lambda_{\text{min}}^{n/2}(\Sigma_n)\sqrt{\lambda_{\text{max}}(\Sigma_n)}}. \end{aligned} \tag{7}$$

From (7),

$$\begin{aligned}
&\frac{\sqrt[n]{M_n}}{\sqrt{n}} \\
&\le \frac{\sqrt[n]{B}}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{n\lambda_{\min}(\Sigma_n)}} \cdot \frac{1}{\sqrt[n]{\lambda_{\max}(\Sigma_n)}} \\
&\to 0 \text{ by hypothesis.}
\end{aligned}$$

Thus the hypotheses of Theorem 1 are satisfied and so the corollary follows.

Proof of Theorem 1:

Step 1.

$$\begin{aligned}
&P_{f_n}(||\underset{\sim}{X}|| \le k) \\
&= \int_{||\underset{\sim}{x}|| \le k} f_n(\underset{\sim}{x}) d\underset{\sim}{x} \\
&\le \int_{||\underset{\sim}{x}|| \le k} g_n(||\underset{\sim}{x}||) d\underset{\sim}{x} \\
&= C(n) \int_0^k g_n(r) r^{n-1} dr, \qquad (8)
\end{aligned}$$

where $C(n) = \frac{2\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2})}$ is the surface area of the n-dimensional unit sphere.

Step 2.

$$\begin{aligned}
&\int_0^k g_n(r) r^{n-1} dr \\
&= \frac{1}{n} \int_0^{k^n} g_n(z^{\frac{1}{n}}) dz \\
&\le \frac{g_n(1)k^n}{n} + \frac{1}{n} \int_0^{k^n} |g_n(z^{\frac{1}{n}}) - g_n(1)| dz \\
&\le \frac{g_n(1)k^n}{n} + \frac{M_n}{n} \int_0^{k^n} |z^{\frac{1}{n}} - 1| dz \\
&= \frac{g_n(1)k^n}{n} + M_n \int_0^k |x-1| x^{n-1} dx \\
&\le \frac{g_n(1)k^n}{n} + M_n \left[\int_0^k x^n dx + \int_0^k x^{n-1} dx \right]
\end{aligned}$$

$$\leq \frac{g_n(1)k^n}{n} + M_n\left(\frac{k^{n+1}}{n+1} + \frac{k^n}{n}\right)$$
$$\leq \frac{g_n(1)k^n}{n} + \frac{2M_n k^{n+1}}{n+1} \tag{9}$$

(as we may assume without loss of generality that $k > 1$ and hence $k \geq \frac{n+1}{n}$ for all large n).

Step 3. By an application of Stirling's approximation,

$$C(n) = O\left(\sqrt{n}\cdot\left(\frac{\sqrt{2\pi e}}{\sqrt{n}}\right)^n\right)$$

Step 4. Hence,

$$\begin{aligned}\frac{C(n)g_n(1)k^n}{n} &= O\left(\frac{g_n(1)}{\sqrt{n}}\cdot\left(\frac{k\sqrt{2\pi e}}{\sqrt{n}}\right)^n\right)\\ &= O\left(\frac{1}{\sqrt{n}}\left(\frac{\sqrt[n]{g_n(1)}}{\sqrt{n}}\cdot k\sqrt{2\pi e}\right)^n\right)\\ &\to 0 \text{ by hypothesis } \underset{\sim}{\text{i}} \text{ of Theorem 1.}\end{aligned}$$

Step 5.

$$\begin{aligned}\frac{C(n)M_n k^{n+1}}{n+1} &= O\left(\frac{M_n}{\sqrt{n}}\cdot\left(\frac{k\sqrt{2\pi e}}{\sqrt{n}}\right)^n\right)\\ &= O\left(\frac{1}{\sqrt{n}}\cdot\left(\frac{\sqrt[n]{M_n}}{\sqrt{n}}\cdot k\sqrt{2\pi e}\right)^n\right)\\ &\to 0 \text{ by hypothesis } \underset{\sim}{\text{ii}} \text{ of Theorem 1.}\end{aligned}$$

Combining Steps 1, 2, 4, and 5, the proof is now complete.

2.3. An Example. We will give a brief example which illustrates the collapse to zero of probabilities of fixed spheres by using multivariate t distributions, a popular alternative to the normal. The formula we give for probabilities of fixed spheres under t distributions could be of some independent use.

Example 1. Suppose $\underset{\sim}{X}$ has the n-dimensional spherically symmetric t distribution with α degrees of freedom ($\alpha > 0$), i.e., $\underset{\sim}{X}$ has the density

$$f(\underset{\sim}{x}) = \frac{\Gamma(\frac{\alpha+n}{2})}{(\alpha\pi)^{\frac{n}{2}}\Gamma(\frac{\alpha}{2})}\ \frac{1}{(1+\frac{\|\underset{\sim}{x}\|^2}{\alpha})^{\frac{\alpha+n}{2}}}. \tag{10}$$

then the probability of a fixed sphere $\{\underset{\sim}{x}\colon ||\underset{\sim}{x}|| \leq k\}$ is

$$\begin{aligned} &P(||\underset{\sim}{X}|| \leq k) \\ &= \frac{C(n)\Gamma\left(\frac{\alpha+n}{2}\right)}{(\alpha\pi)^{\frac{n}{2}}\Gamma(\frac{\alpha}{2})} \int_0^k \frac{r^{n-1}}{(1+\frac{r^2}{\alpha})^{\frac{\alpha+n}{2}}} dr, \end{aligned} \tag{11}$$

where $C(n) = \frac{2\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2})}$ is the surface area of the unit n-sphere. From (11), on transforming to $x = \frac{r^2}{\alpha}$,

$$\begin{aligned} &P(||\underset{\sim}{X}|| \leq k) \\ &= \frac{\Gamma\left(\frac{\alpha+n}{2}\right)}{\Gamma(\frac{\alpha}{2})\Gamma(\frac{n}{2})} \int_0^{\frac{k^2}{\alpha}} \frac{x^{\frac{n}{2}-1}}{(1+x)^{\frac{\alpha+n}{2}}} dx \\ &= \frac{\Gamma\left(\frac{\alpha+n}{2}\right)}{\Gamma(\frac{\alpha}{2})\Gamma(\frac{n}{2})} \left(\frac{k^2}{\alpha}\right)^{\frac{n}{2}} \frac{\Gamma(\frac{n}{2})}{\frac{n}{2}\Gamma(\frac{n}{2})} \cdot 2F_1(\frac{\alpha+n}{2}, \frac{n}{2}; \frac{n}{2}+1; -\frac{k^2}{\alpha}) \end{aligned}$$

(See pp. 284 in Gradshteyn and Ryzhik (1980))

$$= \frac{2k^n}{n\alpha^{\frac{n}{2}}B(\frac{\alpha}{2}, \frac{n}{2})} \cdot {}_2F_1(\frac{\alpha+n}{2}, \frac{n}{2}; \frac{n}{2}+1; -\frac{k^2}{\alpha}) \tag{12}$$

Table 2 uses formula (12) for $k = 1$, with $\alpha = 1, 3, 5, \infty$ and various n. The case $\alpha = \infty$ is given as a basis for comparison (as it corresponds to the normal). From Table 2, it appears that the curse of dimensionality manifests itself in 5 or more dimensions.

Table 2

$P(||\underset{\sim}{X}|| \leq 1)$

$\underline{\alpha}$ (d.f.)

$\underline{n}$ (dimensions)	1	3	5	∞
1	.5	.6090	.6368	.6822
2	.2929	.3505	.3661	.3932
3	.1817	.1955	.1974	.1988
5	.0756	.0577	.0510	.0377
6	.0498	.0308	.0249	.0145
7	.0331	.0163	.0119	.0053
10	.0101	.0024	.0012	.0002

3. Visit Probabilities of Random Sets

Suppose k observations have been sampled at random from the n-dimensional cube. Will any of these points be visible if we restrict attention to a set planted randomly within the cube? For large n, the phenomenon of the curse of dimensionality suggests that in fact the points will probably not be caught inside a randomly placed set. Of course, one has to give a meaning to a set being placed at random within a cube. We look at L_p balls. In our result, we take a random L_p ball by first choosing the center of the ball at random within the cube, and then choosing the radius at random in the admissible interval (to be made precise below).

3.1. Random Balls. We first give our definition of a random L_p ball; there is, of course, a rich theory of random sets.

Definition. Let $\underset{\sim}{\mu}_{n\times 1}$ be chosen according to the uniform distribution in the n dimensional unit cube $[-1,1]^n$ and let r be distributed uniformly in $[0,t]$ where $t = 1 - \max\limits_{1\leq i\leq n} |\mu_i|$. For given $p > 0$, the L_p ball $B = \{\underset{\sim}{x} \colon ||\underset{\sim}{x} - \underset{\sim}{\mu}||_p \leq r\}$ will be called a **Random L_p Ball in the Unit Cube**.

Theorem 2. Let $c(n,p) = \frac{(\Gamma(\frac{1}{p}+1))^n}{\Gamma(\frac{n}{p}+1)}$. Let $\underset{\sim}{X}_1, \underset{\sim}{X}_2, \ldots, \underset{\sim}{X}_k$ be k uniformly distributed points in the unit cube $[-1,1]^n$. Let B be a random L_p ball in the unit cube. Then

$$\begin{aligned} &P(B \text{ does not contain any } \underset{\sim}{X}_i) \\ &= 1 - n! \sum_{j=1}^{k} (-1)^{j+1} \binom{k}{j} \frac{c^j(n,p)\Gamma(nj+1)}{(nj+1)\Gamma(n(j+1)+1)}. \end{aligned} \tag{13}$$

Proof: (13) follows on lengthy and somewhat tedious calculations on using the facts that the volume of an L_p ball of radius r is $\left(\frac{2r}{p}\right)^n \frac{(\Gamma(\frac{1}{p}))^n}{\Gamma(\frac{n}{p}+1)} = (2r)^n c(n,p)$ and that if $\underset{\sim}{\mu}$ is chosen uniformly in $[-1,1]^n$, then the density of $t = 1 - \max\limits_i |\mu_i|$ is nt^{n-1}, $0 \leq t \leq 1$. The case $k = 1$ is theoretically interesting; we state the following corollary.

Corollary 2. The probability that a random L_p ball fails to contain a uniformly distributed point in the unit cube equals $1 - \frac{(n!)^2}{(n+1)(2n)!}c(n,p)$. In particular, for $p = 2$, this probability is $\frac{3}{4}$, $1 - \frac{\pi}{72}$, $1 - \frac{\pi}{480}$, and $1 - \frac{\pi^2}{11200}$ for $n = 1, 2, 3, 4$ respectively.

Remark. The numerical values of $1 - \frac{\pi}{72}$, $1 - \frac{\pi}{480}$ and $1 - \frac{\pi^2}{11200}$ are .9564, .9935 and .9991. Thus the most drastic effect occurs at two dimensions for a single observation. Of course, this should not be interpreted as an occurrence of the curse of dimensionality in two dimensions, as the case of a single observation is only intellectually interesting.

The following table gives the probability that a random sphere will remain unvisited after k observations; it is obtained from (13) by taking $p = 2$. From Table 3, we see that a random sphere in 4 dimensions remains unvisited with 95% probability after 100 observations!

Table 3

P (Random Sphere is unvisited)

	$\underline{k}$ (# observations)				
$\underline{n}$ (dimensions)	1	5	20	50	100
1	.75	.4083	.1736	.0886	.0515
2	.9564	.8437	.6679	.5379	.4447
3	.9935	.9713	.9176	.8580	.8019
4	.9991	.9958	.9854	.9701	.9520
5	.9999	.9995	.9980	.9953	.9916
6	1	.9999	.9998	.9994	.9989
7	1	1	1	.9999	.9999
10	1	1	1	1	1

Table 4

Minimum sample size to have
P (Random Sphere is visited) $> .5$

$\underline{n}$	2	3	4	5
Minimum sample size	70	2400	145,000	1.2×10^7

Table 4 is quite astonishing and calls into question the feasibility of multivariate density estimation beyond 4 dimensions.

4. Sample Sizes Necessary for Full Occupancy

From the results of the preceding sections, one would anticipate that a huge number of observations will be necessary if we desire at least one observation in each member of a partition of a sample space in high dimensions. In this section, we attempt to quantify these sample sizes. These will also be useful for understanding how many dimensions are needed to feel the curse of high dimensions (of course, this is subject to interpretation and cannot be uniquely defined). Note the formal connection to the coupon collector's problem; see Holst (1986) and Kolchin et al (1978).

4.1. One in Each Cell.

If the requirement is at least one in each cell, then one has the classic coupon collector's problem. Given n cells with cell probabilities $p_1, \dots, p_m$, we have the following result. We will use the notation

$$g(\underset{\sim}{p}) = p_1 \cdot \sum_{k=2}^{m} \left\{ \prod_{j=2}^{k} \frac{p_j}{p_j + \cdots + p_m} \right\} \frac{1}{1 - p_1 - \cdots - p_{k-1}}; \tag{14}$$

$$N_{m,r} = \text{ \# samples necessary to have } \geq r \text{ observations in each cell.} \tag{15}$$

Theorem 3. Let $\underset{\sim}{X}_1, \underset{\sim}{X}_2, \dots$ be observations from a distribution and suppose the sample space Ω is partitioned into $S_1, \dots, S_m$ with probabilities $p_1, \dots, p_m$. Then,

$$E(N_{m,1}) = 1 + \sum_{\pi_m} g(\pi_m \underset{\sim}{p}), \tag{16}$$

where $\pi_m \underset{\sim}{p}$ denotes a generic permutation of $\underset{\sim}{p} = (p_1, \dots, p_m)$.

Proof: Write $N_{m,1} = 1 + \omega_2 + \cdots + \omega_m$, where ω_i denotes the waiting time to fill the ith new cell. Suppose the cells are filled in the natural order. Then having filled cells $1, 2, \dots, k-1$, the expected waiting time to fill cell k is $\frac{1}{1-p_1-\cdots-p_{k-1}}$

Now,

$$E(N_{m,1}) = \sum_{\pi_m} E(N_{m,1} \Big| \text{ the cells get filled in the order of } \pi_m \underset{\sim}{p}) \cdot$$
$$P(\text{the cells get filled in the order of } \pi_m \underset{\sim}{p}) \tag{17}$$

and so, (17) follows from (16).

Example 2. We will use formula (12) in Section 2.3 to obtain $E(N_{m,1})$ if the observations are from a multivariate t distribution. For the partition of the sample space we take the annuli $S_i = \{\underset{\sim}{x} : a_{i-1} \leq \|\underset{\sim}{x}\|_2 < a_i\}$, where $a_0 = 0$, $a_m = \infty$, and $a_i = i$ for $0 < i < m$. Thus, using formula (12), for $i < m$,

$$p_i = \frac{2}{n\alpha^{\frac{n}{2}} B(\frac{\alpha}{2}, \frac{n}{2})} \left\{ i^n \cdot {}_2F_1 \left(\frac{\alpha+n}{2}, \frac{n}{2}; \frac{n}{2}+1; -\frac{i^2}{\alpha} \right) - (i-1)^n \cdot \right.$$
$$\left. {}_2F_1 \left(\frac{\alpha+n}{2}, \frac{n}{2}; \frac{n}{2}+1; -\frac{(i-1)^2}{\alpha} \right), \right\} \tag{18}$$

and $p_m = 1 - \sum\limits_{i<m} p_i$.

Substitution of (18) into formula (16) of Theorem 3 leads to the following table. We take $m = 2$ and 3.

Table 5

Average sample size necessary
to fill 2 or 3 annuli

	α (d.f.)					
	1		3		∞	
n (dimensions)	$m = 2$	$m = 3$	$m = 2$	$m = 3$	$m = 2$	$m = 3$
2	3.83	5.92	3.40	5.61	3.19	8.30
3	5.73	7.24	5.36	6.50	5.28	6.92
5	13.31	14.30	17.40	17.88	26.63	26.87
7	30.20	30.94	61.29	61.58	190.73	190.84
10	98.83	99.34	425.25	425.42	5607.95	5608.01

Note the interesting insensitivity to the choice of m!

4.2. Two Cells. The curse of dimensionality, of course, as we see in Table 5, manifests itself with just two cells as well. It is particularly striking when it is desired to have sufficiently many observations in each of the two cells. For instance, one may want to understand certain features of a density within each of two cells and so would like to see a good number of observations in each cell. The following result gives the expected sample size necessary to get r or more observations in one cell and s or more in the other for specified r, s.

Proposition 2. Let $\underset{\sim}{X}_1, \underset{\sim}{X}_2, \ldots$ be independent observations from a density $f(\underset{\sim}{x})$. Let S be any (measurable) subset of the sample space Ω and suppose S has probability $p = 1 - q$ under f. Let $N_{r,s}$ be the minimum number of observations necessary to have r or more observations within S and s or more observations within S^c. Then,

$$E(N_{r,s}) = \frac{r}{p} + \frac{s}{q} - \sum_{i=1}^{s} P(\text{Beta } (r, s-i+1) \geq p) - \sum_{i=1}^{r} P(\text{Beta } (s, r-i+1) \geq q) \quad (19)$$

Corollary 3. The average sample size necessary to have one or more observations within each S and S^c is $\frac{1}{pq} - 1$ and the average sample size necessary to have two or more observations within each S and S^c is $2(\frac{1}{pq} - 1) - 2pq$.

Proof of Corollary 3. The first case corresponds to

$$\begin{aligned} E(N_{1,1}) &= \frac{1}{p} + \frac{1}{q} - P(\text{Beta } (1,1) \geq p) - P(\text{Beta } (1,1) \geq q) \\ &= \frac{1}{pq} - (1-p) - (1-q) \\ &= \frac{1}{pq} - 1. \end{aligned}$$

The second case corresponds to

$$\begin{aligned} E(N_{2,2}) &= \frac{2}{p} + \frac{2}{q} - P(\text{Beta } (2,2) \geq p) - P(\text{Beta } (2,1) \geq p) \\ &\quad - P(\text{Beta } (2,2) \geq q) - P(\text{Beta } (2,1) \geq q) \\ &= \frac{2}{pq} - 6\int_p^1 x(1-x)dx - 2\int_p^1 xdx - 6\int_q^1 x(1-x)dx - 2\int_q^1 xdx, \end{aligned}$$

which simplifies to $2(\frac{1}{pq} - 1) - 2pq$ on easy calculations.

Proof of Proposition 2:

Step 1. Let k be any integer $\geq r + s$. We will find $E(N_{r,s})$ as $r + s + \sum_{k=r+s}^{\infty} P(N_{r,s} > k)$.

Step 2.

$$\begin{aligned} P(N_{r,s} > k) &= P(X < r \text{ or } k - X < s | X \sim \text{ Bin } (k,p)) \\ &= \sum_{j=0}^{r-1} \binom{k}{j} p^j q^{k-j} + \sum_{j=0}^{s-1} \binom{k}{j} q^j p^{k-j} \end{aligned}$$

(since $k \geq r + s$, $X < r$ and $k - X < s$ are disjoint)

$$= \sum_{j=0}^{r-1} \frac{p^j}{j!} \frac{k!}{(k-j)!} q^{k-j} + \sum_{j=0}^{s-1} \frac{q^j}{j!} \frac{k!}{(k-j)!} p^{k-j}. \tag{20}$$

Step 3.

$$\sum_{k=r+s}^{\infty} \frac{k!}{(k-j)!} q^{k-j} = f^{(j)}(q), \text{ where } f(q) = \sum_{k=r+s}^{\infty} q^k = \frac{q^{r+s}}{1-q}.$$

Step 4.

$$\frac{q^{r+s}}{1-q} = \frac{1}{1-q} + \frac{q^{r+s}-1}{1-q} = \frac{1}{1-q} - \sum_{i=1}^{r+s} q^{r+s-i}$$

Hence,

$$f^{(j)}(q) = \frac{j!}{(1-q)^{j+1}} - \sum_{i=1}^{r+s-j} \frac{(r+s-i)!}{(r+s-i-j)!} q^{r+s-i-j}. \tag{21}$$

Step 5. Substitution of (21) into (20) yields

$$\begin{aligned}
&E(N_{r,s}) \\
&= r + s + \sum_{j=0}^{r-1} \frac{p^j}{j!} f^{(j)}(q) + \sum_{j=0}^{s-1} \frac{q^j}{j!} f^{(j)}(p) \\
&= r + s + \sum_{j=0}^{r-1} \frac{p^j}{j!} \left\{ \frac{j!}{p^{j+1}} - \sum_{i=1}^{r+s-j} \frac{(r+s-i)!}{j!} q^{r+s-i-j} \right\} \\
&\quad + \sum_{j=0}^{s-1} \frac{q^j}{j!} \left\{ \frac{j!}{q^{j+1}} - \sum_{i=1}^{r+s-j} \frac{(r+s-i)!}{j!} p^{r+s-i-j} \right\} \\
&= r + s + \frac{r}{p} + \frac{s}{q} - \sum_{i=1}^{r+s} \sum_{j=0}^{\min(r-1,r+s-i)} \binom{r+s-i}{j} p^j q^{r+s-i-j} \\
&\quad - \sum_{i=1}^{r+s} \sum_{j=0}^{\min(s-1,r+s-i)} \binom{r+s-i}{j} q^j p^{r+s-i-j}.
\end{aligned} \tag{22}$$

(by an interchange in the order of summation in the second terms).

Step 6.

$$\sum_{i=1}^{r+s} \sum_{j=0}^{\min(r-1,r+s-i)} \binom{r+s-i}{j} p^j q^{r+s-i-j}$$

$$= \sum_{i=1}^{s} \sum_{j=0}^{r-1} \binom{r+s-i}{j} p^j q^{r+s-i-j} + \sum_{i=s+1}^{r+s} \sum_{j=0}^{r+s-i} \binom{r+s-i}{j} p^j q^{r+s-i-j}$$

$$= r + \sum_{i=1}^{s} P(\text{Bin } (r+s-i, p) \leq r-1)$$

$$= r + \sum_{i=1}^{s} P(\text{Beta } (r, s-i+1) \geq p) \tag{23}$$

(this connection between the Beta and the Binomial distributions is well known).

Similarly, the second double sum in (22) equals

$$s + \sum_{i=1}^{r} P(\text{Beta } (s, r-i+1) \geq q). \tag{24}$$

Step 7. Substitution of (23) and (24) into (22) yields the Proposition.

Corollary 3 and Proposition 2 are used to compute the average sample size necessary to get $\geq r$ observations in each cell $||\underset{\sim}{X}|| \leq 1$ and $||\underset{\sim}{X}|| > 1$ for multivariate t distributions; we use $r = 1, 2, 5, 10$.

Table 6A

Average sample size necessary
to have $\geq r$ observations in 2 cells

	$\underline{\alpha \text{ (d.f.)}}$									
	1					∞				
$\underline{n}$ (dimensions)	$r = 1$	$r = 2$	$r = 5$	$r = 10$	$r = 30$	$r = 1$	$r = 2$	$r = 5$	$r = 10$	$r = 30$
2	3.83	7.24	17.33	34.24	102.43	3.19	5.91	13.63	26.23	76.60
3	5.73	11.15	27.55	55.04	165.11	5.28	10.24	25.20	50.31	150.91
4	8.74	17.28	43.06	86.12	258.40	11.16	22.17	55.33	110.65	331.86
5	13.31	26.48	66.14	132.28	398.83	26.56	53.06	132.63	265.25	795.76
6	20.12	40.15	100.35	200.70	602.41	68.91	137.80	344.50	688.99	2068.97
7	30.25	60.43	151.06	302.12	906.34	172.42	344.83	862.07	1724.14	5660.38
10	99.02	198.02	495.05	990.10		5000	10000	25000	50000	

From Table 6A, we see that if we want merely 10 observations in each of the two annuli, then in 7 dimensions we already need more than 1700 samples just on an average. Practically, by 5 or 6 dimensions, we already start to strongly feel the curse of dimensionality.

On the basis of these calculations, we give a subjective MINIMUM sample size recommendation in various dimensions. Below these recommended sample sizes, it will likely be futile to attempt statistical inference in the corresponding dimension. In specific problems, one can set a goal and give more objective sample sizes.

Table 6B

$\underset{\sim}{n}$ (dimension)	2	3	4	5	6	7
Minimum Recommended Sample Size	110	175	350	800	2100	5700

5. Empty Cells.

Due to the phenomenon of the curse of dimensionality, it is expected that if histogram density estimation is attempted in high dimensions, then there would be an abundance of empty cells. The fixed sample distribution of the number of empty cells is in general extremely complex; there is one case, namely the case of equal cell probabilities, in which it can be written relatively simply. We state it first as a basis for comparison in the subsequent results. The asymptotic distribution of the number of empty cells in histogram density estimation will depend on the growth of the number of cells relative to the sample size and the cells themselves; it is an interesting theoretical question and will be reported in a forthcoming article (DasGupta (1997)). We will use the notation $T_{k,m}$ for the number of cells among m remaining empty after k observations.

5.1. Equiprobable Cells.

The result stated below is distribution free and has nothing to do directly with curse of dimensionality. It will be used as a standard for comparison in Table 7. First we give a definition; see Anderson (1989).

Definition. The Stirling number of second kind, $S(k,r)$, is the number of distinct unordered partitions of a positive integer k into r positive integers $x_1, x_2, \dots, x_r$.

Proposition 3. Let $\underset{\sim}{X}_1, \underset{\sim}{X}_2, \dots, \underset{\sim}{X}_k$ be k independent observations from a density $f(\underset{\sim}{x})$ and let $S_1, S_2, \dots, S_m$ be a partition of the sample space Ω such that $P_f(S_i) = \frac{1}{m}$ for each i. Then

$$P(T_{k,m} = r) = \frac{m!}{r!} \frac{S(k, m-r)}{m^k} \qquad (25)$$

Proof: If there are exactly r empty cells, then the k observations get distributed into

$m - r$ nonempty cells in $(m-r)!S(k, m-r)$ ways and the r empty cells can be chosen in $\binom{m}{r}$ ways and so the Proposition follows.

5.2. General Cells. We will try to get a qualitative understanding of how the curse of high dimensions affects the number of empty cells by considering the important multivariate normal case and the cells $||\underset{\sim}{X}|| \leq 1, \ldots, m-2 \leq ||\underset{\sim}{X}|| \leq m-1$, $||\underset{\sim}{X}|| > m-1$ and by comparing with (25). First we give one theoretical result which will quantify in the most general case the extreme difficulty of getting all nonempty cells.

Proposition 4. Let $\underset{\sim}{X}_1, \underset{\sim}{X}_2, \ldots, \underset{\sim}{X}_k$ be k independent observations from a density $f(\underset{\sim}{x})$ and let $S_1, S_2, \ldots, S_m$ be a partition of the sample space Ω, with $P_f(S_i) = p_i$. Then

$$P(T_{k,m} \geq 1) = \sum_{i=1}^{m-1} (-1)^{i+m-1} \sigma(\underset{\sim}{p}, k, i) \tag{26}$$

where $\sigma(\underset{\sim}{p}, k, i) = \sum_{j_1 < \ldots < j_i} (p_{j_1} + p_{j_2} + \cdots + p_{j_i})^k$.

In particular, if $m = 3$, then

$$\begin{aligned} &P(T_{k,m} \geq 1) \\ &= (p_1 + p_2)^k + (p_2 + p_3)^k + (p_1 + p_3)^k - p_1^k - p_2^k - p_3^k \end{aligned} \tag{27}$$

and if $m = 4$, then

$$\begin{aligned} &P(T_{k,m} \geq 1) \\ &= p_1^k + p_2^k + p_3^k + p_4^k - (p_1 + p_2)^k - (p_1 + p_3)^k - (p_1 + p_4)^k \\ &- (p_2 + p_3)^k - (p_2 + p_4)^k - (p_3 + p_4)^k \\ &+ (p_1 + p_2 + p_3)^k + (p_1 + p_2 + p_4)^k + (p_1 + p_3 + p_4)^k + (p_2 + p_3 + p_4)^k. \end{aligned} \tag{28}$$

Proof: Let A_i be the event $\{T_{k,m} = i\}$. Then $P(T_{k,m} \geq 1) = P(\bigcup_{i=1}^{m-1} A_i)$. (26) will follow by use of the inclusion-exclusion formula and (27), (28) follow from (26).

Example 3. The result in Proposition 4 is used below to compute the probability that at least one cell remains empty after $k = 100$ observations for the multivariate normal case.

The sets $S_1, S_2, \ldots, S_m$ of the partition are annuli as in the previous examples. These probabilities are in Table 7; the case of equiprobable cells is also given for comparison. In Table 8 we report the threshold dimension at which the probability of at least one empty cell is 90% or more.

Table 7

Probability of one or more empty cells
in the normal case

$\underline{m}$ (# annuli)

	2	3	4	2	3	4
		(Equispaced)			(Equiprobable)	
$\underline{n}$ (dimensions)						
2	0	0	.3114	0	0	0
3	0	0	.0456	0	0	0
5	.0214	.0214	.0214	0	0	0
7	.5878	.5878	.5878	0	0	0
10	.9803	.9803	.9762	0	0	0

Table 8

Minimum # dimensions at which
P (at least one empty cell) $\geq .9$

$\underline{m}$ (# equispaced annuli)

	2	3	4
$\underline{n}$	9	9	9

Example 4. Finally, we report some simulations on the distribution of $T_{k,m}$, the number of cells remaining empty after k observations. For equiprobable cells, the distribution is given in (25) in closed form and so no simulations were necessary. We use $k = 100$ and the cells are annuli again.

Table 9

Distribution of the # empty cells
in the normal case

$\underline{m}$ (# equispaced annuli)

	2		3		4	
	$P(T=0)$	$P(T=1)$	$P(T=0)$	$P(T=1)$	$P(T=0)$	$P(T=1)$
$\underline{n}$ (dimensions)						
2	1	0	1	0	.6886	.3114
3	1	0	1	0	.9544	.0456
5	.9786	.0214	.0214	.0214	.9786	.0214
7	.4122	.5878	.5878	.5878	.4122	.5878
10	.0197	.9803	.9803	.9803	.0238	.9721

References

Anderson, I. (1989). A First Course in Combinatorial Mathematics, Oxford University Press,New York.

DasGupta, A. (1997). On the asymptotic distribution of the number of empty cells in histogram density estimates (Forthcoming).

Devroye, L. and Gyorfi, L. (1985). Nonparametric Density Estimation: The L_1 View, John Wiley, New York.

Epanechnikov, V. A. (1969). Nonparametric estimation of a multivariate probability density, *Th. Prob. and Appl.*, **14**, 153–158.

Gradshteyn, I. S. and Ryzhik, I. M. (1980). Table of Integrals, Series, and Products, Academic Press, New York.

Holst, L. (1986). On birthday, collectors', occupancy and other classical urn problems, *Int. Stat. Review*, **54**, 15–27.

Izenman, A. J. (1991). Recent developments in nonparametric density estimation, *Jour. Amer. Stat. Assoc.*, **86**, 205–224.

Kendall, D. G. (1974). Foundations of a theory of random sets, *Stoch. Geometry* (eds. E. F. Harding and D. G. Kendall), 322–376, John Wiley, New York.

Kolchin, V. F., Sevastyanov, B. A., and Chistyakov, V. P. (1978). Random Allocations, John Wiley, New York.

Matheron, G. S. (1975). Random Sets and Integral Geometry, John Wiley, New York.

Nadaraya, E. A. (1983). Nonparametric Estimation of a Probability Density and Regression Curve, Publ. Office of Tbilisi University, Tbilisi, USSR.

Rosenblatt, M.(1956). Remark on some nonparametric estimates of a density function, *Ann. Math. Stat.*, **27**, 832–837.

Scott, D. W. (1992). Multivariate Density Estimation: Theory, Practice, and Visualization, John Wiley, New York.

Scott, D. W. and Wand, M. P. (1991). Feasibility of multivariate density estimates, *Biometrika*, **78**, 197–206.

Silverman, B. W. (1986). Density Estimation for Statistics and Data Analysis, Chapman and Hall, London.

Wegman, E. J. (1990). Hyperdimensional data analysis using parallel coordinates, *Jour. Amer. Stat. Assoc.*, **85**, 664–675.

Wertz, W. (1978). Statistical Density Estimation: A Survey, Vandenhoeck and Ruprecht, Gottingen.