smfd3.tex
Day 3. 16.5.2012

3. LIKELIHOOD RATIO TESTS

We turn now to the general case: *composite* H_0 v. *composite* H_1. We may not be able to find UMP (best) tests. Instead, we seek a general procedure for finding *good* tests.

Let θ be a parameter, H_0 be a null hypothesis – a set of parameter values Θ_0, such that H_0 is true iff $\theta \in \Theta_0$, and similarly for H_1, Θ_1. It is technically more convenient to take H_1 *more general than* H_0, and we can do this by replacing H_1 by "H_1 or H_0". Then $\Theta_0 \subset \Theta_1$.

With L the likelihood, we write

$$L_0 := \sup_{\theta \in \Theta_0} L(\theta), \qquad L_1 := \sup_{\theta \in \Theta_1} L(\theta).$$

As with MLE: the size of L_1 is a measure of how well the data supports H_1. So to test H_0 v. H_1, we use test statistic the *likelihood ratio (LR)* statistic,

$$\lambda := L_0/L_1,$$

and critical region: *reject* H_0 *if* λ *is too small.*

Since $\Theta_0 \subset \Theta 1$, $L_0 \leq L_1$, so

$$0 \leq \lambda \leq 1.$$

In standard examples, we may be able to find the distribution of λ. But in general this is hard to find, and we have to rely instead on large-sample asymptotics.

Theorem (S. S. WILKS, 1938). If θ is a one-dimensional parameter, and λ is the likelihood-ratio statistic for testing $H_0 : \theta = \theta_0$ v. $H_1 :$ θ unrestricted, then under the usual regularity conditions for MLEs (I.3),

$$-2 \log \lambda \to \chi^2(1) \qquad (n \to \infty).$$

Proof. $\lambda = L_0/L_1$, where $L_0 = L(\mathbf{X}; \theta_0)$, $L_1 = L(\mathbf{X}; \hat{\theta})$, with $\hat{\theta}$ the MLE (I.1). So

$$\log \lambda = \ell(\theta_0) - \ell(\hat{\theta}) = \ell_0 - \ell_1,$$

say. But

$$\ell(\theta_0) = \ell(\hat{\theta}) + (\theta_0 - \hat{\theta})\ell'(\hat{\theta}) + \frac{1}{2}(\theta_0 - \hat{\theta})\ell''(\theta^*),$$

with θ^* between θ_0 and $\hat{\theta}$, by Taylor's Theorem. As $\hat{\theta}$ is the MLE, $\ell'(\hat{\theta}) = 0$. So

$$\log\lambda = \ell_0 - \ell_1 = \frac{1}{2}(\theta_0 - \hat{\theta})^2\ell''(\theta^*), \qquad -2\log\lambda = (\theta_0 - \hat{\theta})^2[-\ell''(\theta^*)].$$

By consistency of the MLE (I.3), $\hat{\theta} \to \theta_0$ a.s. as $n \to \infty$. So also $\theta^* \to \theta_0$. So

$$\begin{aligned} -\ell''(\theta^*) &= -\ell''(\mathbf{X};\theta^*) = n.\frac{1}{n}\sum_1^n[-\ell''(X_i;\theta^*)] \\ &\sim nE[-\ell''(X_i;\theta^*)] \qquad (LLN) \\ &= nI(\theta^*) \qquad \text{(definition of information per reading)} \\ &\sim nI(\theta_0) \qquad (\theta^* \to \theta_0). \end{aligned}$$

By I.3,

$$(\hat{\theta} - \theta_0)\sqrt{nI(\theta_0)} \to \Phi, \qquad (\hat{\theta} - \theta_0)^2.nI(\theta_0) \to \Phi^2 = \chi^2(1),$$

using Φ^2 as shorthand for 'the distribution of the square of a standard normal random variable'. So

$$-2\log\lambda \to \chi^2(1). \qquad //$$

Higher Dimensions. If $\theta = (\theta_r, \theta_s)$ is a vector parameter, with

θ_r an r-dimensional parameter of interest,

θ_s an s-dimensional nuisance parameter,

to test $H_0 : \theta_r = \theta_{r,0}$ v. $H_1 :$ θ_r unrestricted. Similar use of the large-sample theory of MLEs for vector parameters (which involves Fisher's *information matrix*) gives

Theorem (Wilks, 1938). Under the usual regularity conditions,

$$-2\log\lambda \to \chi^2(r) \qquad (n \to \infty).$$

Note that the dimensionality s of the nuisance parameter plays no role: what counts is r, the dimension of the parameter of interest (i.e., the difference in dimension between H_1 and H_0). (We think here of a fully specified

parameter, as in H_0, as a point – of dimension 0, and of H_1 of dimension r, like θ_r. There need not be any vector-space structure here. Recall that degrees of freedom (df) correspond to effective sample size, and that for every parameter we estimate we 'use up' one df, so reducing the effective sample size by the number of parameters we estimate, so reducing also the available information. For background, see e.g. [BF], Notes 3.8, 3.9.)

Example: Normal means $N(\mu, \sigma^2)$, σ unknown.

Here μ is the *parameter of interest*, σ is a *nuisance parameter* – a parameter that appears in the *model*, but not in the *hypothesis* we wish to test.

$$H_0: \quad \mu = \mu_0 \qquad v. \quad H_1: \quad \mu \text{ unrestricted.}$$

$$L = \frac{1}{\sigma^n (2\pi)^{n/2}} . \exp\{-\frac{1}{2}\sum_1^n (x_i - \mu)^2/\sigma^2\},$$

$$L_0 = \frac{1}{\sigma^n (2\pi)^{n/2}} . \exp\{-\frac{1}{2}\sum_1^n (x_i - \mu_0)^2/\sigma^2\} = \frac{1}{\sigma^n (2\pi)^{n/2}} . \exp\{-\frac{1}{2} n S_0^2/\sigma^2\},$$

in an obvious notation. The MLEs under H_1 are $\hat{\mu} = \bar{X}$, $\hat{\sigma}^2 = S^2$, as before, and under H_0, we obtain as before $\sigma = S_0$. So

$$L_1 = \frac{e^{-\frac{1}{2}n}}{S^n (2\pi)^{\frac{1}{2}n}}; \qquad L_0 = \frac{e^{-\frac{1}{2}n}}{S_0^n (2\pi)^{\frac{1}{2}n}}.$$

So

$$\lambda := L_0/L_1 = S^n/S_0^n.$$

Now

$$n S_0^2 = \sum_1^n (X_i - \mu_0)^2 = \sum [(X_i - \bar{X}) + (\bar{X} - \mu_0)]^2$$

$$= \sum (X_i - \bar{X})^2 + n(\bar{X} - \mu_0)^2 = n S^2 + n(\bar{X} - \mu_0)^2$$

(as $\sum (X_i - \bar{X}) = 0$):

$$\frac{S_0^2}{S^2} = 1 + \frac{(\bar{X} - \mu_0)^2}{S^2}.$$

But $t := (\bar{X} - \mu_0)\sqrt{n-1}/S$ has the Student t-distribution $t(n-1)$ with n df under H_0, so

$$S_0^2/S^2 = 1 + t^2/(n-1).$$

The LR test is: reject if

$\lambda + (S/S_0)^n$ too small;

$S_0^2/S^2 = 1 + t^2/(n-1)$ too big;

t^2 too big: $|t|$ too big, which is the Student t-test:

The LR test here is the Student t-test.

2. Normal variances $N(\mu, \sigma^2)$, μ unknown (a nuisance parameter). Test

$$H_0: \quad \sigma = \sigma_0 \qquad v. \qquad H_1: \quad \mathcal{S} > \sigma_{\prime}.$$

Under H_0, $\ell = const - n \log \sigma_0 - \frac{1}{2} \sum (X_i - \mu)^2/\sigma_0^2$.
$\partial \ell / \partial \mu = 0$: $\sum_1^n (X_i - \mu) = 0$:

$$\hat{\mu} = \frac{1}{n} \sum_1^n X_i = \bar{X}.$$

So

$$L_0 = \frac{1}{\sigma_0^n (2\pi)^{n/2}} . \exp\{-\frac{1}{2} \sum_1^n (x_i - \mu_0)^2/\sigma_0^2\} = \frac{1}{\sigma_0^n (2\pi)^{n/2}} . \exp\{-\frac{1}{2} n S^2/\sigma_0^2\}.$$

Under H_1, $\ell = const - n \log \sigma - \frac{1}{2} \sum (X_i - \mu)^2/\sigma^2$. As above, the maximising value for μ is $\bar{X}$, and as $\sum_1^n (X_i - \bar{X})^2 = nS^2$,

$$\ell = const - n \log \sigma - \frac{1}{2} \sum (X_i - \mu)^2/\sigma^2 = const - n \log \sigma - \frac{1}{2} n S^2/\sigma^2.$$

$\partial / \partial \sigma = 0$: $-n/\sigma + nS^2/\sigma^3 = 0$: $\sigma^2 = S^2$.

There are two cases: I. $\sigma_0 < S$. II. $\sigma_0 \geq S$.

In Case I, S belongs to the region $\sigma > \sigma_0$ defining H_1, so the maximum over H_1 is attained at S, giving as before

$$L_1 = \frac{e^{-\frac{1}{2}n}}{S^n (2\pi)^{\frac{1}{2}n}}. \quad \text{So} \quad \lambda = \frac{L_0}{L_1} = \frac{S^n}{S_0^n} \exp\Big\{-\frac{1}{2} n \Big[\frac{S^2}{\sigma_0^2} - 1\Big]\Big\}. \qquad (Case\ I).$$

In Case II, the maximum of L is attained at S (L increases up to S, then decreases), so its restricted maximum in the range $\sigma \geq \sigma_0 \geq S$ is attained at σ_0, the nearest point to the overall maximum S. Then

$$L_1 = \frac{1}{\sigma_0^n (2\pi)^{n/2}} . \exp\{-\frac{1}{2} \sum_1^n (x_i - \mu_0)^2/\sigma_0^2\} = L_{)}: \qquad \lambda = L_0/L_1 = 1$$

$$(Case\ II).$$

Comparing, λ is a function of $T := S/\sigma_0$:

$$\lambda = 1 \quad \text{if } T \leq 1 \text{ (Case II)}, \qquad T^n \exp\{-\frac{1}{2}n[T^2 - 1]\} \quad \text{if } T \geq 1 \text{ (Case I)}.$$

Now $f(x) := x^n \exp\{-\frac{1}{2}n[x^2 - 1]\}$ takes its maximum on $(0, \infty)$ at $x = 1$, where it takes the value 1 (check by calculus). So (check by graphing λ against T!) the LR test is:

reject if λ too small, i.e. T too big, i.e. S too big – as expected.

Under H_0, nS^2/σ_0^2 is $\chi^2(n-1)$... If c_α is the upper α-point of $\chi^2(n-1)$, reject if $nS^2/\sigma_0^2 \geq c_\alpha$, i.e., *reject if* $S \geq \sigma_0^2 c_\alpha/n$.

Similarly if H_1 is $\sigma < \sigma_)$ and d_α is the lower α-point: *reject if* $S^2 \leq \sigma_0^2 d_\alpha/n$.

III. NON-PARAMETRICS

1. EMPIRICALS; THE GLIVENKO-CANTELLI THEOREM

The first thing to note about Parametric Statistics is that the parametric model we choose will only ever be approximately right at best. We recall *Box's Dictum* (the English statistician George E. P. BOX (1919 –)): *al models are wrong – some models are useful.* For example: much of Statistics uses a normal model in one form or other. But no real population will ever be exactly normal. And even if it were, when we sampled from it, we would destroy normality, e.g. by the need to *round* data to record it; rounded data is necessarily rational, but a normal distribution takes irrational values a.s.

So we avoid choosing a parametric model, and ask what can be done without it. We sample from an unknown population distribution F. One important tool is the *empirical* (distribution function) F_n of the sample $X_1, \ldots, X_n$. This is the (random!) probability distribution with mass $1/n$ at each of the data points X_i. Writing δ_c for the *Dirac* distribution at c – the probability measure with mass 1 at c, or distribution function of the constant c –

$$F_n := \frac{1}{n} \sum_1^n \delta_{X_i}.$$

The next result is sometimes called the *Fundamental Theorem of Statistics.* It says that, in the limit, we can recover the population distribution from the sample: *the sample determines the population in the limit.* It is due to V. I. GLIVENKO (1897-1940) and F. P. CANTELLI (1906-1985), both in 1933, and is a uniform version of Kolmogorov's Strong Law of Large Numbers (SLLN, or just LLN), also of 1933.

Theorem (Glivenko-Cantelli Theorem, 1933).

$$\sup_x |F_n(x) - F(x)| \to 0 \qquad (n \to \infty) \qquad a.s.$$

Proof. Think of obtaining a value $\leq x$ as Bernoulli trials, with parameter (= success probability) $p := P(X \leq x) = F(x)$. So by SLLN, for each fixed x,

$$F_n(x) \to F(X) \qquad a.s.,$$

as $F_n(x)$ is the proportion of successes. Now fix a finite partition $-\infty = x_1 < x_2 < \ldots < x_m = +\infty$. By monotonicity of F and F_n,

$$\sup_x |F_n(x) - F(x)| \leq \max_k |F_n(x_k) - F(x_k)| + \max_k |F(x_{k+1} - F(x_k)|.$$

Letting $n \to \infty$ and refining the partition indefinitely, we get

$$\limsup_n \sup_x |F_n(x) - F(x)| \leq \sup_x \Delta F(x) \qquad a.s.,$$

where $\Delta F(x)$ denotes the jump of F (if any – there are at most countably many jumps!) at x. This proves the result when F is continuous.

In the general case, we use the Probability Integral Transformation (PIT, IS, I). Let $U_1, \ldots, U_n \ldots$ be iid uniforms, $U_n \sim U(0,1)$. Let $Y_n := g(U_n)$, where $g(t) := \sup\{x : F(x) < t\}$. By PIT, $Y_n \leq x$ iff $U_n \leq F(x)$, so the Y_n are iid with law F, like the X_n, so wlog take $Y_n = X_n$. Writing G_n for the empiricals of the U_n,

$$F_n = G_n(F).$$

Writing A for the range (set of values) of F,

$$\sup_x |F_n(x) - F(x)| = \sup_{t \in A} |G_n(t) - t| \leq \sup_{[0,1]} |G_n(t) - t|, \to 0 \qquad a.s.,$$

by the result (proved above) for the continuous case. //

If F is continuous, then the argument above shows that

$$\Delta_n := \sup_x |F_n(x) - F(x)|$$

is *independent* of F, in which case we may take $F = U(0,1)$, and then

$$\Delta_n = \sup_{t \in (0,1)} |F_n(t) - t|.$$

Here Δ_n is the *Kolmogorov-Smirnov (KS) statistic*, which by above is *distribution-free* if F is continuous. It turns out that there is a uniform CLT corresponding to the uniform LLN given by the Glivenko-Cantelli Theorem: $\Delta_n \to 0$ at rate $\sqrt{n}$. The limit distribution is known – it is the *Kolmogorov-Smirnov (KS) distribution*

$$1 - 2\sum_1^\infty (-)^{k+1} e^{-2k^2x^2} \qquad (x \geq 0).$$

It turns out also that, although this result is a limit theorem for *random variables*, it follows as a special case of a limit theorem for *stochastic processes*. Writing B for Brownian motion, B_0 for the Brownian bridge ($B_0(t) := B(t) - t$, $t \in [0,1]$),

$$Z_n := \sqrt{n}(G_n(t) - t) \to B_0(t), \qquad t \in [0,1]$$

(*Donsker's Theorem*: Monroe D. DONSKER (1925-1991), originally, the *Erdös-Kac-Donsker Invariance Principle*). The relevant mathematics here is *weak convergence of probability measures* (under an appropriate topology). Thus, the KS distribution is that of the supremum of Brownian bridge. For background, see e.g. Kallenberg Ch. 14.
Higher dimensions.

In one dimension, the half-lines $(-\infty, x]$ form the obvious class of sets to use – e.g., by differencing they give us the half-open intervals $(a, b]$, and we know from Measure Theory that these suffice. In higher dimensions, obvious analogues are the half-spaces, orthants (sets of the form $\prod_{k=1}^{n}(-\infty, x_k]$), etc. – the geometry of Euclidean space is much richer in higher dimensions. We call a class of sets a *Glivenko-Cantelli class* if a uniform LLN holds for it, a *Donsder class* if a uniform CLT holds for it. For background, see e.g.
[vdVW] A. W. van der VAART & J. A. WELLNER, *Weak convergence and empirical processes, with applications to statistics*, Springer, 1996, Ch. 2.
This book also contains a good treatment of the *delta method* in this context – the *von Mises calculus* (Richard von MISES (1883-1953), or *infinite-dimensional delta method*.

Variants on the problem above include:
1. *The two-sample Kolmogorov-Smirnov test*.

Given two populations, with unknown distributions F, G, we wish to test whether they are the same, on the basis of empiricals F_n, G_m.
2. *Kolmogorov-Smirnov tests with parameters estimated from the data*.

A common case here is *testing for normality*. In one dimension, our hypothesis of interest is whether or not $F \in \{N(\mu, \sigma^2) : \mu \in \mathbf{R}, \sigma > 0\}$. Here (μ, σ) are *nuisance parameters*: they occur in the formulation of the problem, but not in the hypothesis of interest.

Although the Glivenko-Cantelli Theorem is useful, it does not tell us, say, whether or not the law F is absolutely continuous, discrete etc. For, there are discrete G arbitrarily close to an absolutely continuous F (discretise), and absolutely continuous F arbitrarily close to a discrete F (by smooth approximation to F at its jump points). So sampling alone cannot tell us what *type* of law F is – absolutely continuous (with density f, say), discrete, continuous singular, or some mixture of these. So it makes sense for the statistician to *choose* what kind of population distribution he is going to assume. Often (usually), this will be absolutely continuous; again, it makes sense to *assume* what smoothness properties of the density f we will assume. This leads on to the important subject of density estimation, to which we now turn.

2. CURVE AND SURFACE FITTING.

We begin with some background. Suppose we have n points (x_i, y_i), with the x_i distinct, and we wish to *interpolate* them – find a function f with $f(x_i) = y_i$, $i = 1, \ldots, n$. One can of course do this by linear interpolation – just draw a line between each adjacent pair of points, obtaining a continuous piecewise-linear function – but this is not smooth enough for many purposes. One might guess that as a polynomial of degree $n-1$ contains n degrees of freedom (its n coefficients), it might be possible to interpolate by such a polynomial, and this is indeed so (Lagrangian interpolation, or Newtonian divided-difference interpolation). With the x_i equally spaced, there is a whole subject here – the Calculus of Finite Differences (the discrete analogue of the ordinary ('infinitesimal') calculus).

The degree n may be large (should be large – the more data, the better). But, polynomials of large degree are very oscillatory and numerically unstable. We should and do avoid them. One way to do this is to use *splines*. These are continuous functions, which are polynomials of some chosen low degree (*cubic splines* are the usual choice in Statistics) *between* certain special points, called *knots* (or *nodes*), across which the function and as many derivatives as possible are continuous. So a cubic spline is piecewise cubic, with its values and those of its first two derivatives continuous across the knots.

Another relevant piece of background is the *histogram*, familiar from elementary Statistics courses. One represents discrete data diagrammatically, with vertical bars representing how many data points fall in a given subinterval.

Computer implementation is necessary to use methods of this kind in practice. For a general account using the computer language S (from which R, and the proprietary package S-Plus, are derived), see e.g.
W. N. VENABLES & B. D. RIPLEY, *Modern applied statistics with S*, 4th ed., Springer, 2002, 5.6.

Roughness penalty. Using polynomials of high degree, we can fit the data exactly. But we don't do this, because the resulting function would be too rough ('too wiggly'). It is better to fit the data approximately rather than exactly, but obtain a nice smooth function at the end. One way to formalise this (due to I. J. Good (1916-2009) and his pupil R. A. Gaskins in 1971) is to

use a *roughness penalty* – to measure the roughness of the function by some integrated measure – $\int (f'')^2$ is the usual one for use with cubic splines – and minimise a combination of this and the relevant sum of squares (see Ch. V):

$$\min \quad \sum_1^n (y_i - f(x_i))^2 + \lambda^2 \int (f'')^2.$$

Here λ^2 is the *smoothing parameter*. It is under the control of the statistician, who can decide for himself from context how much weight to give to goodness of fit (the first term) and how much to smoothness (roughness being measured by the second term).

1. Density estimation. Suppose we want to find as good a fit to the data as possible using a density function with smoothness properties that we have chosen (see above). One way to do this is to make two key choices:
(a) the *kernel* $K(.)$. This is a density with the required smoothness properties;
(b) the *bandwidth* $h > 0$ (also called the *window width*).
One then defines the *kernel density estimator*

$$\hat{f}(x) := \frac{1}{nh} \sum_1^n K\Big(\frac{x - X_i)}{h}\Big).$$

This is again a density, with the same smoothness properties as K. It turns out that the properties of $\hat{f}$ are mainly determined by h, and the choice of K is less important. We must refer for detail here to a specialised text, e.g.
B. W. SILVERMAN, *Density estimation for statistics and data analysis*, Chapman & Hall, 1986;
R. A. TAPIA & J. R. THOMPSON, *Nonparametric probability density estimation*, Johns Hopkins University Press, 1978.
Such books contain graphics, showing how the kernel density estimates compare with the histograms of the data.

2. Non-parametric regression. We discuss parametric regression in Ch. V below. The ideas above can be used to extend these ideas to a non-parametric setting, using roughness penalties, cubic splines etc. We refer for detail to, e.g., [BF], 9.2.

3. Volatility surfaces. The volatility σ in the Black-Scholes formula is unknown, and has to be estimated – either as *historic volatility* from time-series

data (Ch. VI), or as *implied volatility* – the Black-Scholes price is (continuous and) increasing in σ ('options like volatility'), so one can infer 'what the market thinks σ is' from the prices at which options currently trade. Closer examination reveals that the volatility is not constant, but varies – e.g., with the strike price ('volatility smiles'). Volatility is observed to vary so unpredictably that it makes sense to model is as a stochastic process (*stochastic volatility, SV*). Market data is discrete, but for visual effect it is better to use computer graphics and a continuous representation of such *volatility surfaces*. For a monograph treratment, see
Jim GATHERAL, *The volatility surface: A practitioner's guide*, Wiley, 2006.
Note. The VIX – volatility index (colloquially called the 'fear index') is widely used, and is the underlying for volatility derivatives. It has even affected literature (see e.g. John Harris' novel *The fear index*, Hutchinson, 2011).

3. NON-PARAMETRIC LIKELIHOOD

At first glance, 'non-parametric likelihood' seems a contradiction in terms (an oxymoron – 'square circle', etc.) But it turns out that maximum-likelihood estimation (MLE) can indeed be usefully combined with non-parametrics. First, we interpret the empirical F_n as a non-parametric MLE (NPMLE) for the unknown true distribution F. For, if the data is $\{x_1, \ldots, x_n\}$, the *likelihood* of F is $L(F) := \prod_1^n \Delta F(x_i)$ (where $\Delta F(x) := F(x) - F(x-)$ is the probability mass on x), $F(\{x\})$). It makes sense to restrict attention to distributions F with support in $\{x_1, \ldots, x_n\}$, that is, absolutely continuous wrt the empirical F_n: $F << F_n$, and F_n does indeed maximise the likelihood over these F (Kiefer & Wolfowitz, 1956). Then it makes sense to call $T(F_n)$ a NPMLE for $T(F)$, where T is some functional – the mean, for example.

Let $X, X_1, \ldots, X_n \ldots$ be iid random p-vectors, with mean $EX = \mu$ and covariance matrix Σ of rank q. In higher dimensions, the distribution function, $P(. \leq .)$, which leads to *confidence intervals*, is replaced by $P(. \in .)$, which leads to *confidence regions* (which covers the unknown parameter with some probability); convexity is a desirable property of such confidence regions. For $r \in (0, 1)$, let

$$C_{r,n} := \{\int X dF : F << F_n, L(F)/L(F_n) \geq r\}.$$

Then $C_{r,n}$ is a convex set, and

$$P(\mu \in C_{r,n}) \to P(\chi^2(q) \leq -2 \log r) \qquad (n \to \infty)$$

(the rate is $O(1/\sqrt{n})$ if $E[\|X\|^4] < \infty$). This is a non-parametric analogue of Wilks' Theorem (II.3 above) (A. Owen 1990; P. Hall 1990): "$-2\log LR \sim \chi^2(q)$". For a monograph account, see
A. OWEN, *Empirical likelihood*, Chapman and Hall, 2001.

In view of results of this type, it is common practice, when we want the distribution of $T(F)$ when F is unknown, to use $T(F_n)$ as an approximation for it. This is commonly known as a *plug-in estimator* (just plug it in as an approximation when we need the exact answer but do not know it); 'empirical estimator', or 'NPMLE', would also be reasonable names.

Suppose we want to estimate an unknown density f, which is known to be *decreasing* on $[0, \infty)$(example: the exponential). A density is the derivative of a distribution; a concave function has a decreasing derivative (when differentiable). The NPMLE f_n of such a density is the (left-hand) derivative of the *least concave majorant* of F_n (Grenander, 1956). This example is interesting in that a CLT is known for it, but with an unusual rate of convergence – *cube-root asymptotics*:

$$n^{1/3}(f_n(t) - f(t)) \to |4f'(t)f(t)|^{1/3}\text{argmax}_h(B(h) - h^2),$$

with B BM and argmax the argument (= point) at which the maximum is attained (Kim and Pollard 1990).
Semi-parametrics.

Consider a multidimensional density

$$f(\mathbf{x}) = const.g(Q(\mathbf{x})), \qquad Q(\mathbf{x}) = (\mathbf{x} - \mu)^T\Sigma^{-1}(\mathbf{x} - \mu).$$

Here $g : \mathbf{R}_+ \to \mathbf{R}_+$ is a function, the *density generator*, to be estimated. This is the *non-parametric* part of the model; (μ, Σ) is as above, the *parametric* part of the model. The model as a whole is then called *semi-parametric*.

Such models are very suited to financial applications. They have been extensively studied; see e.g.
[BKRW] P. J. BICKEL, C. A. J. KLAASSEN, Y. RITOV and J. A. WELLNER, *Efficient and adaptive estimation for semiparametric models*, Springer, 1998.
It turns out that in some cases, ignorance of one part of the model imposes no loss of efficiency when estimating the other part. This is the case for the elliptic model above, essentially for reasons to do with invariance under the action of the affine group. See [BKRW], 4.2.3, 6.3.9, 7.2.4, 7.8.3 for the theory, [BFK] for some applications.