Laws of Large Numbers, Summarizing distributions, models and the likelihood principle

September 19, 2013

The SLLN

If $\{X_t\}$ is i.i.d. for $t = 1, \ldots, \infty$ and $E[X_t]$ is well defined, then

$$\frac{1}{T}\sum_{t=1}^{\infty} X_t = \bar{X}_T \xrightarrow[T\to\infty]{} E[X_t]$$

with probability one. Sometimes this is written as $\bar{X}_T \xrightarrow{\text{a.s.}} E[X_t]$, or as "$\bar{X}_T$ converges almost surely to $E[X_t]$". The i.i.d. distributions of the X_t's imply a probability distribution over all sequences of real numbers $\{x_t\}$, $t = 1, \ldots, \infty$. The SLLN says that the class of all sequences of real numbers that don't converge to $E[X_t]$ has zero probability. Proving this requires a technical argument, so we won't do it. This is a "strong" law of large numbers

The WLLN

It is also true under the same assumptions that

$$P[\left|\bar{X}_T - E[X_t]\right| > \varepsilon] \xrightarrow[T\to\infty]{} 0 \text{ for any } \varepsilon > 0\,.$$

This is a "weak" law of large numbers. It is also sometimes stated as

$$\bar{X}_T \xrightarrow[T\to\infty]{\mathsf{P}} E[X_t]$$

or as "$\bar{X}_T$ converges in probability to $E[X_t]$". Convergence in probability is implied by almost sure convergence, which is why this latter result is called "weak".

Why bother with the WLLN? Both it and the SLLN can be proved with less restrictive assumptions than the i.i.d. assumption we have used here, and WLLN's can be proved, naturally, with less restrictive assumptions than those needed for a SLLN.

Should the SLLN make you feel safe about simulating?

- No matter what $E[X_t]$ is, $\bar{X}_T$ converges to it. Does this imply that it is safe and sensible to choose a big T, form $\bar{X}_T$, and act as if $\bar{X}_T = E[X_t]$?

- Not necessarily. Suppose that we knew the distribution of X_t made $P[X_t = 0] = 1 - \varepsilon$, $P[X_t = \varepsilon^{-2}] = \varepsilon$. Then $E[X_t] = 1/\varepsilon$. This distribution has a well defined expectation, so the SLLN applies.

- But if ε is very small, we will see very many $X_t = 0$ observations before we see any ε^{-2} observations. In fact the more zeros we see in a row, the more convinced we should be that ε must be small and therefore $E[X_t]$ big. Yet of course so long as we see nothing but zeros, $\bar{X}_T$ remains stuck at zero.

- Furthermore, all the standard checks for "convergence" would indicate increasing confidence that the X_T value is close to the truth as the number of sequential zeros increases.

The sample cdf

- Estimating $E[f(X)]$ by taking the sample average of a sequence of i.i.d. draws from the distribution of X is a special case of the following.

- Define the **sample cdf** of the sample $\{x_1, \ldots, x_T\}$ as

$$F_{XT}(a) = \frac{\text{number of } t\text{'s such that } x_t \leq a}{T} .$$

  The sample cdf implies a distribution that puts probability $1/T$ on each value of x_t, $i = 1 \ldots T$.

- If we are interested in some function of the distribution of X, say

$$P[\sin(X) > 0] = \int \mathbf{1}_{\{\sin(x)>0\}} \, dP(x) \, ,$$

  estimate it instead as

$$\int \mathbf{1}_{\{\sin(x)>0\}} \, dP_T(x) \, ,$$

  where P_T is the probability distribution implied by the sample cdf.

- This is equivalent to

$$\frac{1}{T} \sum_{t=1}^{T} \mathbf{1}_{\{\sin(x_t)>0\}}$$

Convergence of the sample cdf, convergence in distribution

- At every point, a sample cdf formed from an i.i.d. sample converges a.s. to F_X, i.e. $F_{XT}(a) \xrightarrow{a.s.} F_X(a)$ at every a.

- Follows from SLLN, because $F_X(a) = E[\mathbf{1}_{\{x \leq a\}}]$ and the sample cdf is the sample average of $\mathbf{1}_{\{x \leq a\}}(X_t)$.

- This is a special case of **convergence in distribution**. A sequence of distribution functions $\{F_T\}$ converges in distribution to the limit F_∞ if and only if $F_T(a) \to F_\infty(a)$ at every a at which F is continuous.

- Equivalently, for every bounded, continuous function f of X, $E[f(X_T)] \to E[f(X_\infty)]$, where X_T is any random variable with the cdf F_T. Sometimes written $X_T \xrightarrow{\mathcal{D}} X_\infty$.

- So long as we stick to expectations of bounded continuous functions (and their monotone limits), the strategy of substituting the sample cdf for the true one to obtain estimates is justified (as much as it can be) by a SLLN.

- Caution: There are interesting f's that don't satisfy these conditions — e.g., the number of local maxima in the pdf.

- $X_T \xrightarrow{a.s.} X_\infty \Rightarrow X_T \xrightarrow{P} X_\infty \Rightarrow X_T \xrightarrow{\mathcal{D}} X_\infty$.

- This is important to know, but treacherous. a.s. convergence makes an assertion about entire random sequences, convergence in probability makes assertions about the pairwise joint distributions of the X_T's with X_∞, and convergence in distribution makes assertions about the univariate distribution functions of the X's. $X_T \xrightarrow{\mathcal{D}} X_\infty$ does not imply that the realized values of X_T and $X\infty$ have to get close to each other.

Quantiles

- The α quantile of a distribution with cdf F is the value of x such that $F(x) = \alpha$.

- The α quantile of the sample cdf is easy to compute: sort the data in the sample, so that $x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \cdots \leq x_{(T)}$. (This notation, in which $x_{(j)}$ is the j'th element in the sorted sample, is standard. $x_{(j)}$ is sometimes called the j'th **order statistic** of the sample.) Then the α sample quantile is $x_{(j)}$, where $(j-1)/T < \alpha T < j/T$, unless $x_{(j)} = \alpha T$ for some j, in which case it is any number between $x_{(j)}$ and $x_{(j+1)}$.

- When F_X is strictly increasing, the corresponding quantile function satisfies $q_X(\alpha) = F^{-1}(\alpha)$.

- The convergence results for the sample cdf therefore provide some assurance that sample quantiles will converge for an i.i.d. sample, but this has to be qualified. If F_X is flat anywhere, i.e. if there are intervals of nonzero length with zero probability, the quantiles corresponding to the endpoints of that interval are undefined, and the corresponding sample quantiles will not converge.

- Even if there are only intervals with low, but still positive, pdf values, estimates of quantiles falling in those intervals will converge very slowly.

Shortest probability intervals

- These may be the best general way to summarize the shape of a pdf or cdf.

- The minimum length set with probability α can be found from the pdf (if it exists) by choosing a set of the form $S_\alpha = \{x \,|\, p(x) \geq \theta\}$ that satisfies $P[S_\alpha] = \alpha$. This set will be a single interval if the pdf has a single local maximum, but otherwise may consist of disconnected segments. A set like this provides an indication of what are high-probability regions.

- Unlike quantiles shortest sets generalize directly to higher dimensions. They are also easier to grasp intuitively than high-dimensional cdf's.

- When densities exist, the value of the pdf at the boundaries of these sets are constant.

Topographical maps

- For a pdf over two dimensions, minimum-area sets of given probability, if we collect a number of them with different probability values, provide a topographical map of the density function.

- In R, you can get a plot of estimated contours of a 2-d sample by use of `bkde2D()` and `contour()`. The former generates a grid of values in $\mathbb{R}^2$ and estimates of the density function at those points. It uses a "kernel" method to estimate the density —

$$\hat{p}(\vec{x}) = \sum_j k(\vec{x}_j - \vec{x}) \, .$$

  The kernel might be, e.g. proportional to a normal density function. The `contour()` function uses the output of `bkde2D()` to produce the plot.

Covariance matrices

- For an $n \times 1$ random vector X, $\Sigma = E[(X - EX)(X - EX)']$ is the covariance matrix. Its diagonal elements are the variances of the individual random variables in the X vector. Its i, j'th off-diagonal element, $\sigma_{ij} = E[(X_i - EX_i)(X_j - EX_j)] = \mathrm{Cov}(X_i, X_j)$, is the **covariance** of X_i with X_j.

- $|\mathrm{Cov}(X_i, X_j)| \leq \sqrt{\mathrm{Var}(X_i)\,\mathrm{Var}(X_j)}$.

- We define the **correlation** of X_i with X_j as

$$\rho(X_i, X_j) = \mathrm{Cov}(X_i, X_j) / \sqrt{\mathrm{Var}(X_i)\,\mathrm{Var}(X_j)}\,.$$

- If $\rho(X_i, X_j) = \pm 1$, then X_i is an exact linear function of X_j.

- X_i and X_j pairwise independent $\quad \Rightarrow \quad \rho(X_i, X_j) = 0$.

- The reverse impllication is *not* true in general.

- It is natural to take ρ to be a measure of how strongly related two variables are. This is reasonable when the joint distribution of the variables is a lot like a joint normal distribution (which we have yet to define) — having a pdf with a single, round, peak and dropping off rapidly for large values of X. But it can be a poor measure in other cases. [E.g., what is the correlation of X with X^2 when the pdf of X is symmetric around zero?]

Analyzing covariance and correlation matrices

- Obviously the diagonal elements of Σ, being variances of individual random variables, are a measure of the "spread" of their distributions, as in the univariate case.

- The off-diagonal elements of the covariance matrix are a measure of the strength of pairwise relations among the variables.

- But there can be strong multivariate relations among variables that don't show up in pairwise correlations.

Characteristics of covariance matrices

- A matrix Σ is **positive semi-definite** (p.s.d.) if and only if for every conformable vector c, $c'\Sigma c \geq 0$.

- The covariance matrix Σ of a random vector X must be p.s.d. because (as you should be able to verify for yourself) $c'\Sigma c = \mathrm{Var}(c'X)$, and a variance cannot be negative.

- Note also that Σ is symmetric, meaning $\Sigma = \Sigma'$. This follows from the fact that $\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i)$.

- Any symmetric, p.s.d. matrix can be a covariance matrix.

Eigenvalue decomposition

- Any symmetric, p.s.d., $n \times n$ matrix Σ can be decomposed as

$$\Sigma = \sum_{i=1}^{n} v_i \lambda_i v_i' = V \Lambda V' \,,$$

  where $\lambda_i,\ i = 1, \ldots, n$ are non-negative real numbers, v_i are $n \times 1$ vectors, V is a $n \times n$ matrix with the v_i as columns, and Λ is a diagonal matrix with the λ_i down the diagonal.

- V satisfies $V'V = I$, which is to say that V is an **orthonormal** matrix. The columns of V are the right **eigenvectors** of Σ (because $\Sigma V = V\Lambda$ and the λ_i are the **eigenvalues**.

- Matlab, R or Rats will find V and Λ for you with a single command.

From eigenvalue decomposition of Σ to components of X

- The X vector can be respresented as

$$X = \sum_{i=1}^{n} \lambda_i z_i v_i \ ,$$

  where z_i, $i = 1, \ldots, n$ are i.i.d. with mean zero and variance 1.

- The z_i, or sometimes the vector random variables $\lambda_i z_i v_i$, are known as the **principal components** of X. The v_i's associated with large λ_i's correspond to directions in $\mathbb{R}^n$ in which the X vector varies a lot, while those with small λ_i's correspond to directions with very little variation.

- We could calculate principal components of the correlation matrix also. Even though the correlation matrix is just a rescaling of the Σ matrix, its principal components will be different. That is, the correlation matrix is $D^{-\frac{1}{2}}\Sigma D^{-\frac{1}{2}}$, where D is a diagonal matrix with $\sqrt{\mathrm{Var}(X_i)}$ on the diagonal — this is what we mean by saying the correlation matrix is a rescaling of the covariance matrix. But if we take the eigenvalue decomposition $R = WMW'$ of the correlation matrix R, we will find $V \neq D^{\frac{1}{2}}W$ and there is no simple correspondence between the eigenvalues of R and Σ.

- This reflects a general fact about principal component decompositions — they are not scale invariant. Change the units of measurement of some of the components of X and you will change the principle components decomposition.

- Sensitivity to scaling is not the only pitfall to look out for in using the results of a principal components analysis. If a large number of closely related variables are added to the X vector, the first principal component will eventually reflect mainly the common component of those variables.

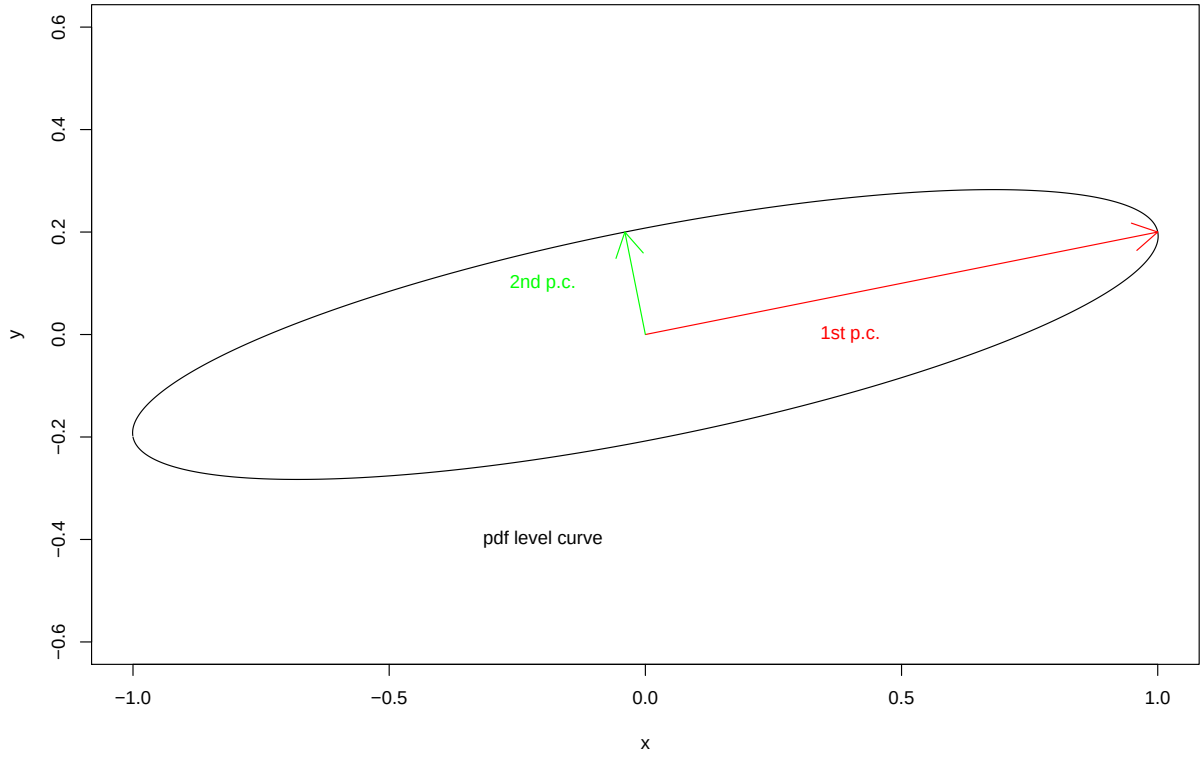
- This might be desired behavior. But often we are tempted to use p.c. analysis when we have several imperfect measures of two or more concepts and we are looking for relations between the concepts. For example we have three measures of education and 3 of income, each of them imperfect. Principal components on this vector will give a different answer with the full 6-dimensional X than what we get if we leave out any element of the X vector.

- Nonetheless principal components decompositions of Σ and/or R are useful descriptive devicesin many cases.

- Not in all cases. As with any function of a distribution we might use to summarize its shape, whether the summary is useful or not depends on the class of distributions we have in mind and what our uncertainties about the distribution are. As with variances, covariance matrices may not exist. Even when they do exist, they may be misleading as measures of spread or of dependence between variables.

2-dimensional geometric interpretation

- If the joint pdf of X, Y has same-shaped elliptical level curves centered at zero, i.e. if the pdf can be written as $p(ax^2+bxy+y^2)$ with $b^2 < 4ac$, or equivalently as $p([x,y]M[x,y]')$ with M positive definite (meaning p.s.d. but with all eigenvalues strictly positive), then if X, Y have a finite covariance matrix, it is proportional to M^{-1}, the eigenvectors of the covariance matrix (and of M) are the principal axes of the ellipses, and the lengths of the principal axes are proportional to the square roots of the eigenvalues of the covariance matrix.


2nd p.c.
1st p.c.
pdf level curve
y
x
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−1.0
−0.5
0.0
0.5
1.0