Convexity points in linear regression

Richard M. Dudley

Massachusetts Institute of Technology

This is joint work with Xia Hua. Consider the simple linear regression model $Y_j = a + bx_j + \varepsilon_j$ where $x_1 < x_2 < \cdots < x_n$ are non-random design points, Y_j are also observed, the "errors" ε_j are unobserved i.i.d. $N(0, \sigma^2)$, $n \geq 2$, and σ^2 is unknown. It's well known that a and b can be uniquely estimated by least squares or equivalently by maximum likelihood, giving $\hat{a}$ and $\hat{b}$. We then observe the residuals $r_j = Y_j - \hat{a} - \hat{b}x_j$. For numbers $s_1, ..., s_n$ which may be either $\varepsilon_1, ..., \varepsilon_n$ or $r_1, ..., r_n$, say that there is a *turning point* at $j = 2, ..., n-1$ if $(s_{j-1} - s_j)(s_{j+1} - s_j) > 0$, or a *convexity point* if $(s_j - s_{j-1})/(x_j - x_{j-1}) < (s_{j+1} - s_j)/(x_{j+1} - x_j)$. There is a convexity point in the errors at j if and only if there is one in the residuals. That is not true for turning points but it becomes approximately true for large n. Thus the indicators t_j of having a turning point at j are weakly dependent. The indicators c_j of having a convexity point at j in the errors and thus in the residuals are 2-dependent. Suppose for simplicity that the spacings $x_j - x_{j-1}$ are all equal. The variables $X_j = c_{2j} + c_{2j+1} - 1$ are 1-dependent and symmetric. The talk will focus on the distribution of the total number of convexity points.