1 Transformations

One typical feature of the data sets in HEP is that they are truncated, that is the data is bounded. In statistics we say the data has finite support. For univariate data the boundary consists simply of the two endpoints of an interval. For Dalitz plots the boundary is given by the kinematics of the decay.

Why is this a problem? Unfortunately nonparametric density estimators have a difficult time dealing with boundaries. The reason for this is as follows. Consider the following artificial data set, which was generated by a flat distribution on the interval $[1,2]$. In figure 1 the observations are drawn as stars along the x-axis. Assume we want to estimate the density at the point $x1$. We have also drawn the kernel function in the graph, centered at $x1$. In the computation of the nonparametric kernel density estimator we sum over all the observations, weighted by the height of the kernel function. Next consider the point $x2$. Again we sum up all the weights, but now there are hardly any observation to the right of $x2$, so the value of $\widehat{f}$ at $x2$ will be smaller than at $x1$ also it obviously should be the same. In figure 2 we draw the kernel density estimate of the same data set for several values of h, together with the true density.It is clear that the density estimates underestimate the true density near the boundary and, because they have to integrate to 1, they overestimate the density in the middle.

A number of ways have been studied to deal with the boundary problem. The most satisfactory one for our purposes is to simply eliminate the boundary by transforming the data to a new set of variables whose support is the whole real line.

We will first show how this works for univariate data. Say the original observations are $X_1,..,X_n$ and they are located in the interval $[A,B]$. Then we will find new observations $Y_1,..,Y_n$ with $Y_k = H(X_k)$; $k = 1,..,n$. The transformation H can be any function that is monotonically increasing, continuous and that maps $A \to -\infty$ and $B \to \infty$. We will do this in two steps. First we map the interval $[A,B]$ onto $[0,1]$ by $Y_k^* = \frac{X_k - A}{B-A}$. Now we need a function that maps the unit interval to the whole real line. Actually, as statisticians we know a whole class of functions that do the opposite, that is map the real line to the unit interval, namely any cumulative distribution function (CDF) of a random variable defined on the whole real line. Say Z is any continuous random variable with $P(a < Z < b) > 0$ for any $a < b$ and Z has density ϕ. Then the cumulative distribution function Φ of Z is defined by

$$\Phi(x) = \int_{-\infty}^{x} \varphi(t)dt \tag{1}$$

Now $\Phi : (-\infty,\infty) \to [0,1]$, therefore its inverse function Φ^{-1}, if it exists, has $\Phi^{-1} : [0,1] \to (-\infty,\infty)$. Therefore any cumulative distribution function of this kind will serve for our purpose. Specifically we will use the most common of these CDF's, the standard normal CDF which is given by

$$\Phi(x) = \int_{-\infty}^{x} \varphi(t)dt = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt \tag{2}$$

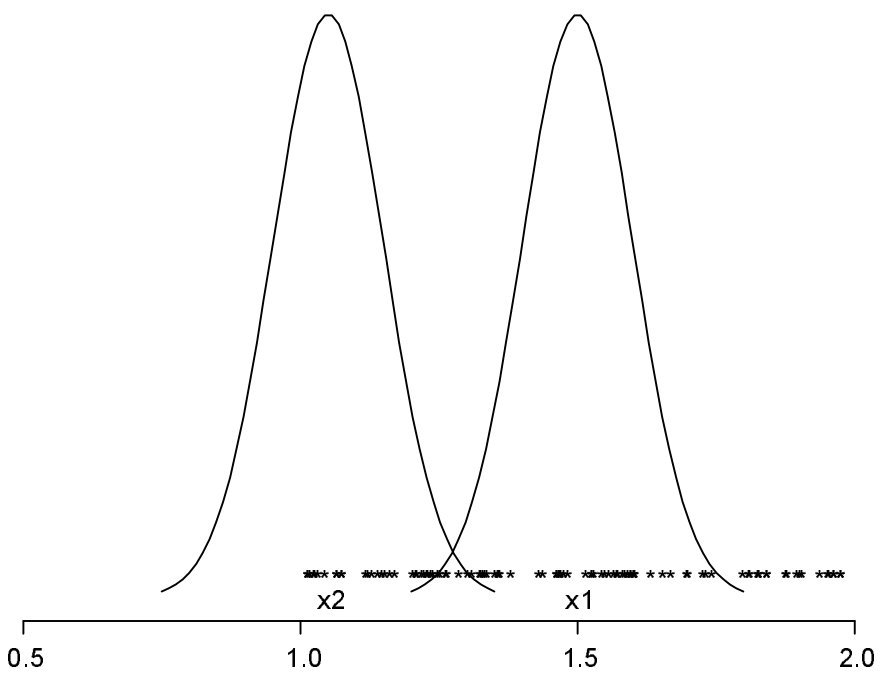

Figure 1:

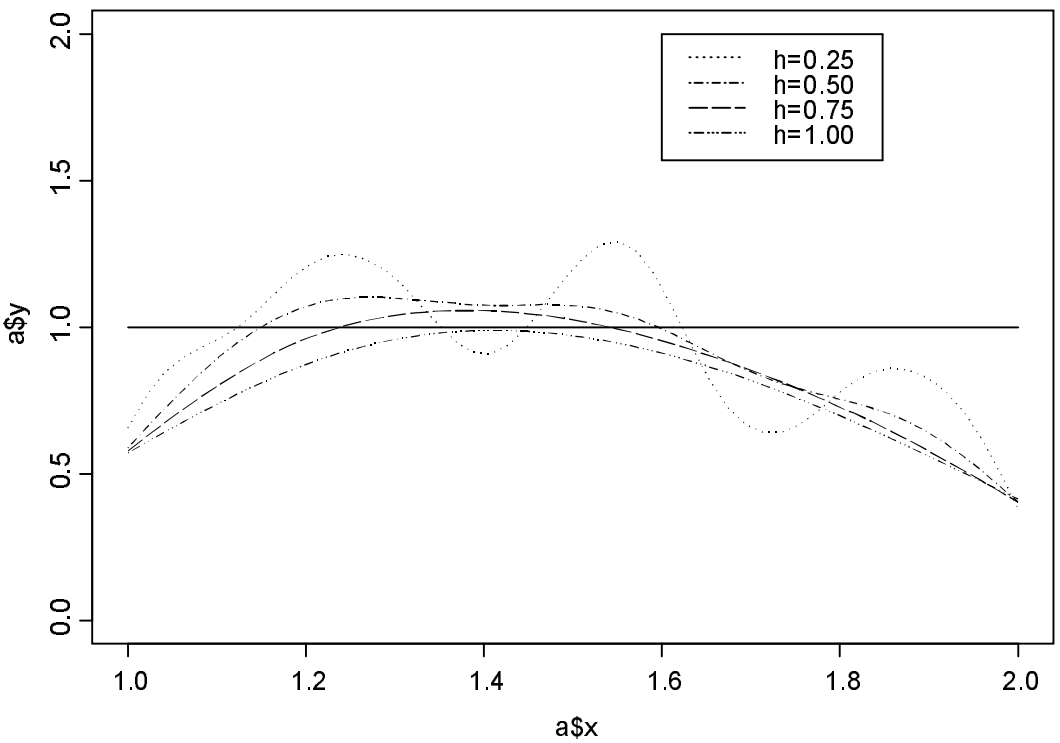

Figure 2:

Then our transformation H is given by

$$Y_k = H(X_k) = \Phi^{-1}\left(\frac{X_k - A}{B - A}\right) \quad k = 1, .., n \tag{3}$$

In figure 3 we have histograms of another data set from a flat distribution on $[1,2]$, this time with 1000 events. In the left pane is the histogram for the original data and in the right pane for the transformed data. Clearly the transformed data set has no boundary.

When we transform the data, what happens to the density of the original data and to its parameters, which are of course the objects of interest? Say the original density was given by $f(x;\theta)$ where θ is the vector of parameters. Let's denote the density of the transformed data by g. Then

$$P\left(Y < y\right) = P\left(H(X) < y\right)$$

$$= P\left(X < H^{-1}(y)\right) = \int_{A}^{H^{-1}(y)} f(x;\theta)dx \tag{4}$$

and so

$$g(y;\theta) = \frac{d}{dy}P\left(Y < y\right) = f(H^{-1}(y);\theta) \cdot \frac{d}{dy}H^{-1}(y) \tag{5}$$

Now if we use the standard normal density for the transformations we have

$$H^{-1}(y) = A + (B - A) \cdot \Phi(y) \text{ and } \frac{d}{dy}H^{-1}(y) = (B - A) \cdot \varphi(y) \tag{6}$$

and 5 becomes

$$g(y;\theta) = f(A + (B - A) \cdot \Phi(y);\theta) \cdot (B - A) \cdot \varphi(y) \tag{7}$$

We can now find the parameter estimates $\widehat{\theta}$ of θ via maximum likelihood. The log-likelihood function $l(\theta)$ is given by

$$l(\theta) = -2\sum_{i=1}^{n} \log g(Y_i;\theta) = -2\sum_{i=1}^{n} \log f(X_i;\theta) \cdot (B - A) \cdot \varphi(Y_i) \tag{8}$$

and we can see that it has a rather simple form.

How does this generalize to Dalitz plots? Because of the way in which the boundary is described for Dalitz plots this turns out to be quite straight forward. Consider the following Dalitz plot for the $D^+ \to kk\pi$ decay, together with the boundary drawn using $M = 1.8693$, $m_1 = 0.493677$, $m_2 = 0.493677$, $m_3 = 0.13956995$.Here the boundary is as follows: $m_{12}^2 \in [A, B]$ where $A = (m_1 + m_2)^2$ and $B = (M - m_3)^2$, and $m_{23}^2 \in \left[L(m_{12}^2), U(m_{12}^2)\right]$ for some known functions

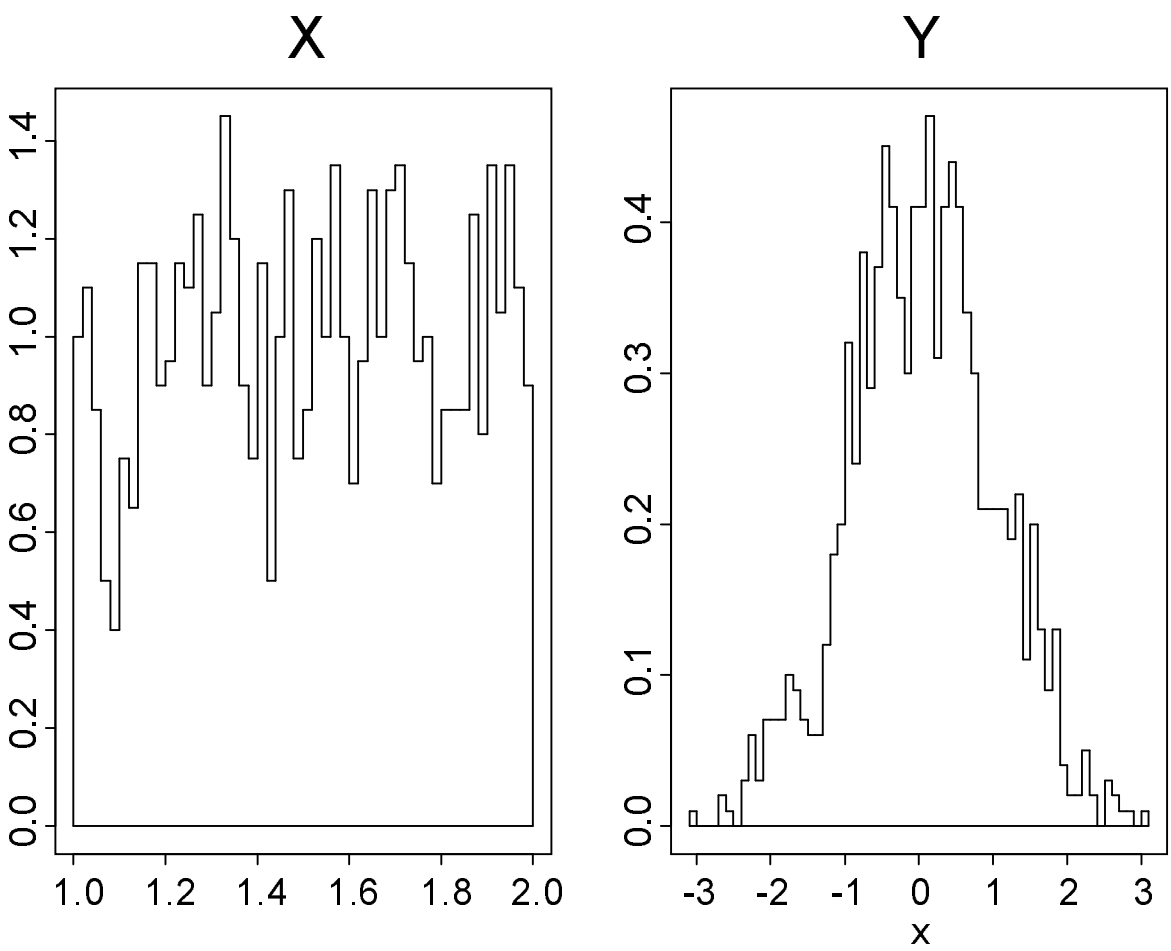

Figure 3:

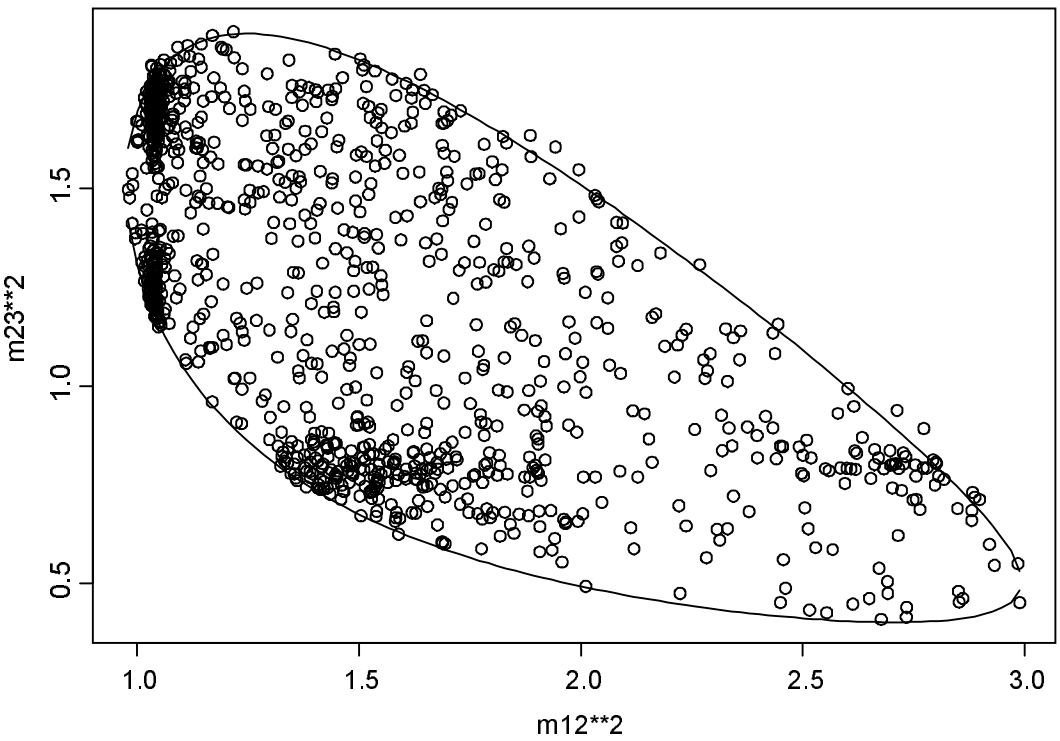

Figure 4:

L and U. Say the observations in the Dalitz plot are $(X_1, Y_1), .., (X_n, Y_n)$. The we can find the transformed observations $(U_1, V_1), .., (U_n, V_n)$ by

$$\begin{aligned} U_i &= h_1(X_i) = \Theta^{-1}\left(\tfrac{X_i - A}{B-A}\right) \\ V_i &= h_2(Y_i) = \Theta^{-1}\left(\tfrac{Y_i - L(X_i)}{U(X_i) - L(X_i)}\right) \end{aligned} \tag{9}$$

Again the question is, what is the density of the transformed variables? Again denoting the density of the transformed random variable by g we get from the general formula for multivariable transformations

$$g(u, v; \theta) = f(k_1(u), k_2(v); \theta) \cdot |\mathbf{J}| \tag{10}$$

where (k_1, k_2) is the inverse transform of $H = (h_1, h_2)$ and is given by

$$\begin{aligned} k_1(u) &= A + (B - A) \cdot \Theta(u) \\ k_2(v) &= L(u) + (U(u) - L(u)) \cdot \Theta(v) \end{aligned} \tag{11}$$

and $\mathbf{J}$ is the Jacobian matrix of the inverse transform, that is

$$\mathbf{J} = \det \begin{vmatrix} \frac{\delta k_1}{\delta u} & \frac{\delta k_1}{\delta v} \\ \frac{\delta k_2}{\delta u} & \frac{\delta k_2}{\delta v} \end{vmatrix} = (A + (B - A) \cdot \varphi(u)) \cdot (L(u) + (U(u) - L(u)) \cdot \varphi(v)) \tag{12}$$

Again we can find the parameter estimates $\widehat{\theta}$ of θ via maximum likelihood. The log-likelihood function $l(\theta)$ is given by

$$l(\theta) = -2 \sum_{i=1}^{n} \log f(X_i, Y_i; \theta) \cdot (B - A) \cdot \varphi(U_i) \cdot (U(X_i) - L(X_i)) \cdot \varphi(V_i) \tag{13}$$

In figure 5 we draw the scatterplot for the transformed data set for the $D^+ \to kk\pi$ decay. As we can see, the data now has no boundaries but the structure in the data is still present.

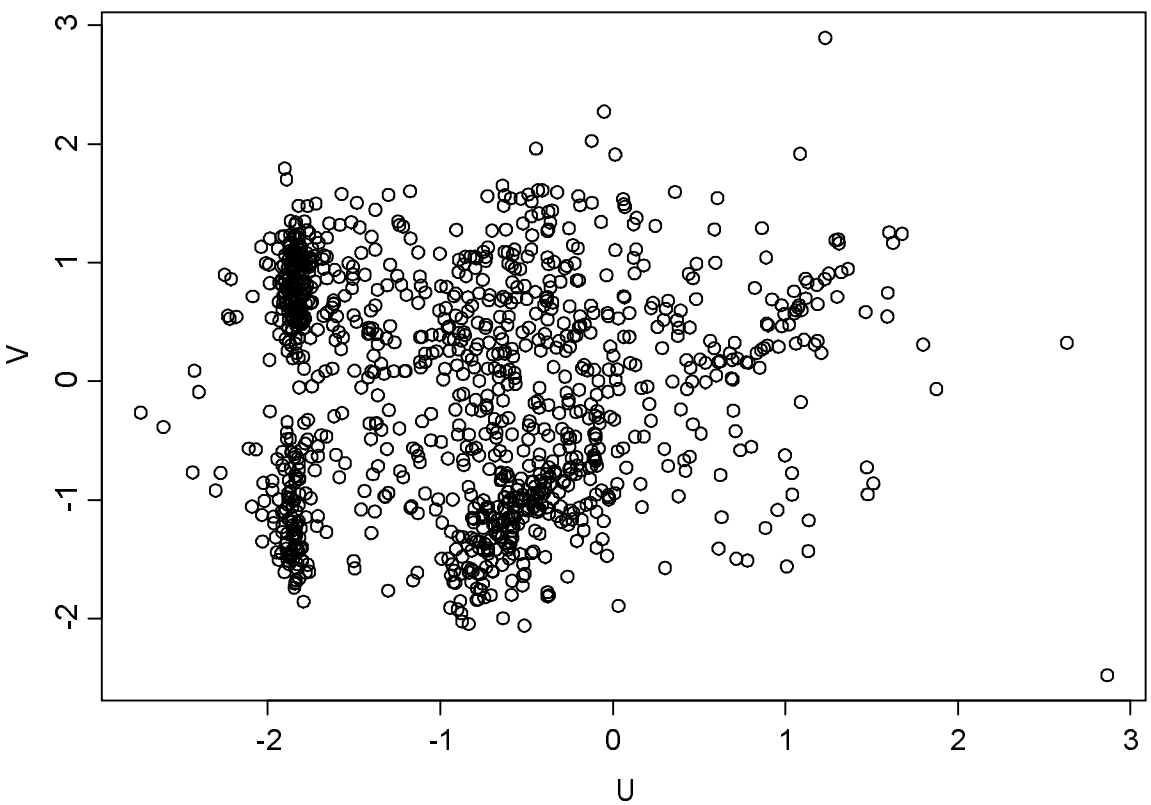

Figure 5: