Copula Fitting (with kernel smoothing). Note by Dr Richard Diamond, CQF.

Canonical estimation operates with pseudo-samples U rather than original data X, consists of five columns of historical data, used for estimation of correlation matrix $\mathbf{\Sigma}_{5\times5}$. Once chosen how to convert the historical data into uniform the **sampling from copula** algorithm implementation is a straightforward by-step. Choices are **converting to Normal variable by *differencing*** $\Delta\mathbf{X} \to \mathbf{Z}$ or **applying probability integral transform and *kernel smoothing*** $\mathbf{X} \to \mathbf{U} \to \mathbf{Z}$.

For Gaussian copula, we estimate linear correlation ρ on Normally distributed $\mathbf{Z}$ so $\mathbf{\Sigma} = \rho(\mathbf{Z})$, for t copula we estimate correlation on the ranks of $\mathbf{X}$ so $\mathbf{\Sigma}_S = \rho(\mathbf{U})$ for Spearman's rho, while separate formula $\mathbf{\Sigma}_\tau = \rho_\tau(\mathbf{X})$ is defined for Kendall's tau. To convert into linear correlation $\rho = 2\sin\left(\frac{\pi}{6}\rho_S\right)$ and $\rho = \sin\left(\frac{\pi}{2}\rho_\tau\right)$ elementwise. This converted matrix is not guaranteed to be positive definite as required for Cholesky – so the nearest correlation matrix is obtained.

Historical sample data $\mathbf{X}^{\text{Hist}}$ (five columns credit spreads/default probabilities/hazard rates) is converted to pseudo-samples $\mathbf{U}^{\text{Hist}} = \widehat{F}(\mathbf{X}^{\text{Hist}})$. That is achieved by special transformation of data by its own *Empirical CDF* and involves *kernel density estimation* in order to guarantee uniformity.[10] Estimation done without making assumption about distribution of marginals is non-parametric and called Canonical Maximum Likelihood.[11]

Let's consider notation as we go from copula fitting (ie, calibration) to simulation,

- $\mathbf{Z} = \Phi^{-1}(\mathbf{U})$ obtained from pseudo-samples, so can be expressed as $\mathbf{Z}^{\text{Hist}}$. Use $\mathbf{\Sigma} = \rho(\mathbf{Z}^{\text{Hist}})$ The shortcut which avoids kernel smoothing is first, take differences $\mathbf{X} = \Delta\mathbf{X}^{\text{Hist}}$ and second, standartise $\mathbf{Z}_t^{(j)} = \frac{\mathbf{X}_t^{(j)} - \mu_j}{\sigma_j}$ for each row (observation) t of column j.
- For calculation of copula density, $\mathbf{U}_t \equiv \mathbf{U}_{t,1\times5}$ refers to **a row** of values for five reference names as observed at time t.
- $\mathbf{Z_{t+}}^{\text{Sim}}$ or simply $\mathbf{Z_{t+}}$ is a vector of simulated 1×5 Standard Normal random variables, and so $\mathbf{U_{t+}} = \Phi(\mathbf{Z_{t+}})$ for Gaussian or $\mathbf{U_{t+}} = T_\nu(\mathbf{Z_{t+}})$ for t copula.

  For the simulated 1×5 $\mathbf{U_{t+}}$, each value is converted to default time $u \to \tau$ using *its own* term structure of hazard rates

$$\tau \sim Exp(\hat{\lambda}_{1Y}, \dots, \hat{\lambda}_{5Y}).$$

Important Disclaimer. Elliptical copulae might fail to fit dependence structure of empirical data (eg, higher density of tail observations, low density of the middle high-peaked observations). That is a model risk the copula method. A quick recipe is to **check bivariate scatters between the columns of U** – the scatter should have the familiar pattern of Elliptical copula density.

[10]There is no analytical formula for Empirical CDF function. It is obtained via a set of algorithms.

[11]Each column of $\mathbf{X}^{\text{Hist}}$ is 'a marginal' with its own univariate distribution that is usually bi-modal for raw credit spreads. Therefore, we have to work with *changes* in spreads $\Delta\mathbf{X}$ (daily or weekly).

Kernel Smoothing

The term refers to the estimation (fitting) of analytical probability density function $\widehat{f}()$ to the data. Most software-implemented kernel smoothers fit probability density function (PDF), from which additional steps have to be taken to obtain CDF $\widehat{F}$ – those are numerical integration over kernel PDF and interpolation. Altogether, the set of algorithms is known as Probability Integral Transform:

$$\mathbf{U} = \widehat{F}(\mathbf{X})$$

Performing MLE on pseudo-samples $\mathbf{U}$ instead of the original data $\mathbf{X}$ is a superior approach. For example, applying the familiar linear correlation formula on ranks $\mathbf{U}$ immediately delivers Spearman's rho, a rank correlation measure.

$$\mathbf{\Sigma}_S = \rho(\mathbf{U})$$

- $\mathbf{X} \to \mathbf{U} \to \mathbf{Z}$. Kernel smoothing on $\mathbf{X}^{\text{Hist}}$ by Empirical CDF algorithm (as implemented in Matlab/R/NAG functions), where implementation guarantees the uniformity of $\mathbf{U}^{\text{Hist}}$.
- $\mathbf{X} \to \mathbf{Z} \ldots \mathbf{U}$. Hidden assumption that original data $\mathbf{X}^{\text{Hist}}$ is converted to near-Normal, for example, by differencing $\mathbf{X} = \Delta\mathbf{X}^*$. Next steps are standardization $\mathbf{Z}_t^{(j)} = \frac{\mathbf{X}_t^{(j)} - \mu_j}{\sigma_j}$ and inferring pseudo-samples $\mathbf{U} = \Phi(\mathbf{Z})$. However, empirical pseudo-samples obtained this way (without kernel smoothing) might be insufficiently uniform.[12]

Where possible *use the ready implementation of kernel smoothing that gives Empirical CDF*, such as Matlab *ksdensity(), and check the uniformity of the output* $\mathbf{U}$ by plotting a histogram for each column (reference name).

When using kernel smoothing **the choice of bandwidth is very important**! Think of it as a bucket of observations for cumulative probability step (standard deviation on uniform scale). MATLAB's *ksdensity()* calibrates some optimal bandwidth, however you might be able to get better smoothing result in terms of uniformity of output $\mathbf{U}$ by interactively experimenting with the 'bw' setting from default down to circa 0.0001.

NAG kernel density estimation (PDF only) does require bandwidth as a ready input, so interactive experiment is necessary. Setting the bandwidth (window width) too high results in the data being represented as fully Normal and therefore, *oversmoothed* and highly correlated across names. Setting the bandwidth too low represents data very close to original (*undersmoothed*) and results in u_i that are zero or close, creating a problem with TINV calculation $T_\nu^{-1}(\mathbf{U})$.

- Each data column of $\mathbf{X}^{\text{Hist}}$ might require calibration of its own bandwidth setting. Here, for kernel PDF data are *changes* in credit spreads/default probabilities/hazard rates.
- Credit monitors rely on weekly changes and drop $1 - 3\%$ of extreme observations. [13]

[12] $\mathbf{Z} \ldots \mathbf{U}$ by $\mathbf{U} = \Phi(\mathbf{Z})$ is actually **the wrong way** but we are trying to see what $\mathbf{U}$ is implied.

[13] In utmost generality, one can look at changes between 5Y hazard rates averaged per period.