Flexible Density Estimation: an Introduction

Feng Li

feng.li@cufe.edu.cn

School of Statistics and Mathematics
Central University of Finance and Economics

Flexible density estimation

↛ Introduction

- Density estimation consecrates on modeling the relationship between the response $\mathbf{y}$ with covariates $\mathbf{x}$ with flexible density function $\mathrm{f}(\cdot)$

$$\mathbf{y} = \mathrm{f}(\mathbf{x}, \theta)$$

  flexible: the density feature θ are modeled in a flexible way.

- An example: GLM: density estimation with flexible mean function $\eta(\mu) = \mathbf{X}\boldsymbol{\beta}$ via the linkage.
- Two main factors that influence the efficiency of the density estimation,
  (1) choice of flexible densities, and
  (2) ways of constructing densities features.

Flexible density estimation

↛ Univariate or multivariate

- (Relevantly) simpler in univariate response
  - Mixture of experts
  - Nonparametric methods: kernel regression, splines...
- More tricky in the multivariate case
  - Flexible multivariate density is difficult to construct *per se*
  - Not only modeling the density features in each marginal model
  - But also multivariate correlations and other dependences need to take into account.

Mixture distributions

- For a given x, a mixture distribution $p(y|x)$ is a finite mixture

$$\sum_{k=1}^{K} \omega_k f_k(y_i|\theta_k), \ i = 1, ..., n.$$

- Latent variable formulation for MCMC

$$\Pr(s_i = k) = \omega_k$$
$$y_i|(s_i = k) \sim f_k(y_i|\theta_i)$$

- Two-block Gibbs sampler
  - Sample $s = (s_1, ..., s_n)$ conditional on $(\theta_1, ..., \theta_k)$.
  - Sample each θ_k conditional on the allocation s.
- A smooth mixture model is a finite mixture density with weights that are smooth function of the covariates, e.g

$$\omega_k(x) = \frac{\exp(x'\gamma_k)}{\sum_{r=1}^{K} \exp(x'\gamma_r)}$$

ME, SMR and SAGM models

- Mixture-of-Experts (ME) (Jacobs *et al.* (1991))
  - A mixture of regressions where the mixing probabilities are functions of covariates.
  - Flexibly model the mean regression and frequently used in the machine learning literature.
  - The components are often linear homoscedastic regressions or even constant functions.
  - *simple-and-many* approach.
- Smoothly Mixing Regression (SMR) (Geweke & Keane (2007))
  - A generalization of the ME model for regression density estimation
  - Fail to fit heteroscedastic data even with a very large number of components
- Smooth Adaptive Gaussian Mixtures (SAGM) (Villani *et al.* (2008))
  - A smooth finite mixture of Gaussian densities with the mixing probabilities.
  - The **mixing probabilities**, the **components means** and **components variances** modeled as functions of the covariates.
  - Bayesian variable selection are in all three sets of covariates.
  - *complex-but-few* approach — Enough flexibility is used within the mixture components so that the number of components can be kept to

Smooth mixture of asymmetric student's t densities

↛ The model

- The split-t density is

$$c \cdot \kappa(\mu, \phi, \nu) I(y \leqslant \mu) + c \cdot \kappa(\mu, \lambda\phi, \nu) I(y > \mu),$$

where $\kappa(\mu, \phi, \nu) = \left(\frac{\nu}{\nu + \frac{(y-\mu)^2}{\phi^2}} \right)^{(\nu+1)/2}$ is the kernel of student t density and c is the normalization constant.

- Each of the four parameters μ, ϕ, λ and ν are connected to covariates as

$$\begin{aligned} \mu &= \beta_{\mu 0} + x_t' \beta_\mu \\ \ln \phi &= \beta_{\phi 0} + x_t' \beta_\phi \\ \ln \lambda &= \beta_{\lambda 0} + x_t' \beta_\lambda \\ \ln \nu &= \beta_{\nu 0} + x_t' \beta_\nu \end{aligned}$$

but any smooth link function can equally well be used in the MCMC methodology.

- This make it possible e.g. to have the degrees of freedom smoothly

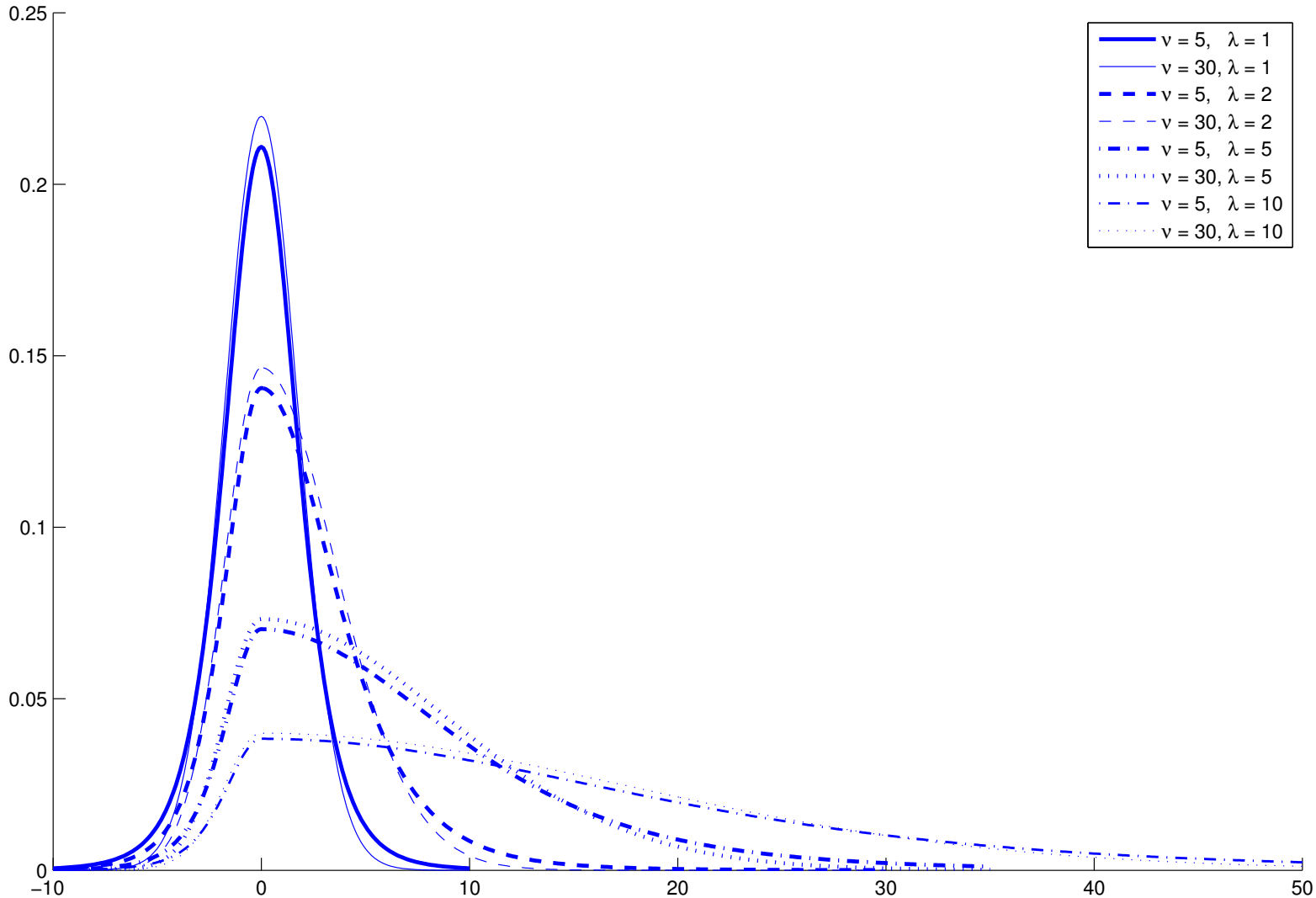


Figure: Graphical display of the split-t density with location parameter $\mu = 0$ and scale parameter $\lambda = 1.8$.

Smooth mixture of asymmetric student's t densities

↛ Discussion — Why not over-fit?

- Variable selection (details in next page)
  - Automatically reduce the model's complexity.
  - Investigate the importance of covariates.
  - More efficient.
- Automatically add components to make each component simpler.
- Evaluating the out-of-sample log predictive density score(LPDS) – details in "model comparison" .

Smooth mixture of asymmetric student's t densities

↛ Model comparison

- Why not marginal likelihood?
  - The key quantity is Bayesian model comparison is the marginal likelihood.
  - The marginal likelihood is sensitive to the choice of prior, which is especially true when the prior is not very informative (Kass, 1993).
- We use B-fold cross-validation of the log predictive density score(LPDS)
  - $\mathrm{B}^{-1} \sum_{\mathrm{b}=1}^{\mathrm{B}} \ln \mathrm{p}\left(\tilde{\mathrm{y}}_{\mathrm{b}} | \tilde{\mathrm{y}}_{-\mathrm{b}}, \mathrm{x}\right)$
  - Compute the LPDS for ME, SMR, SAGM and our split model with different components.
  - Compare the differences of LPDS.

Flexible regression models

↛ Introduction

- Flexible models of the regression function $E(y|x)$ has been an active research field for decades.
- Attention has shifted from kernel regression methods to spline-based models.
- Splines are regression models with flexible mean functions.
- Example: a simple spline regression with only one explanatory variable with truncated linear basis function can be like this

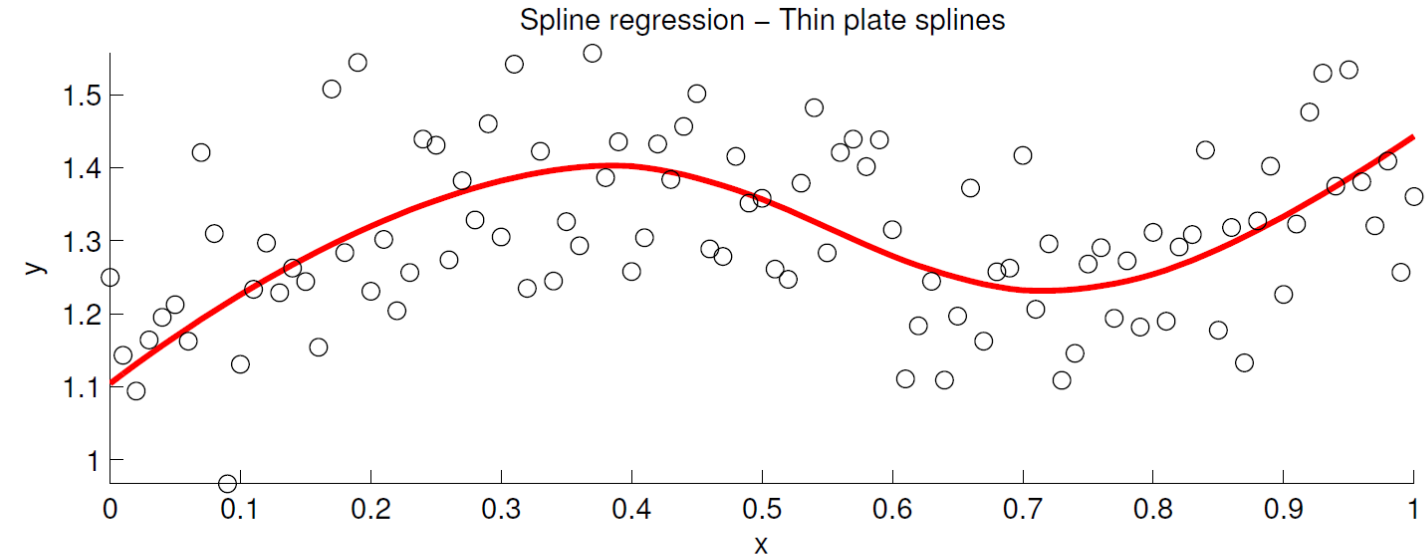
$$y = \alpha_0 + \alpha_1 x + \beta_1 (x - \xi_1)_+ + ... + \beta_q (x - \xi_q)_+ + \varepsilon$$

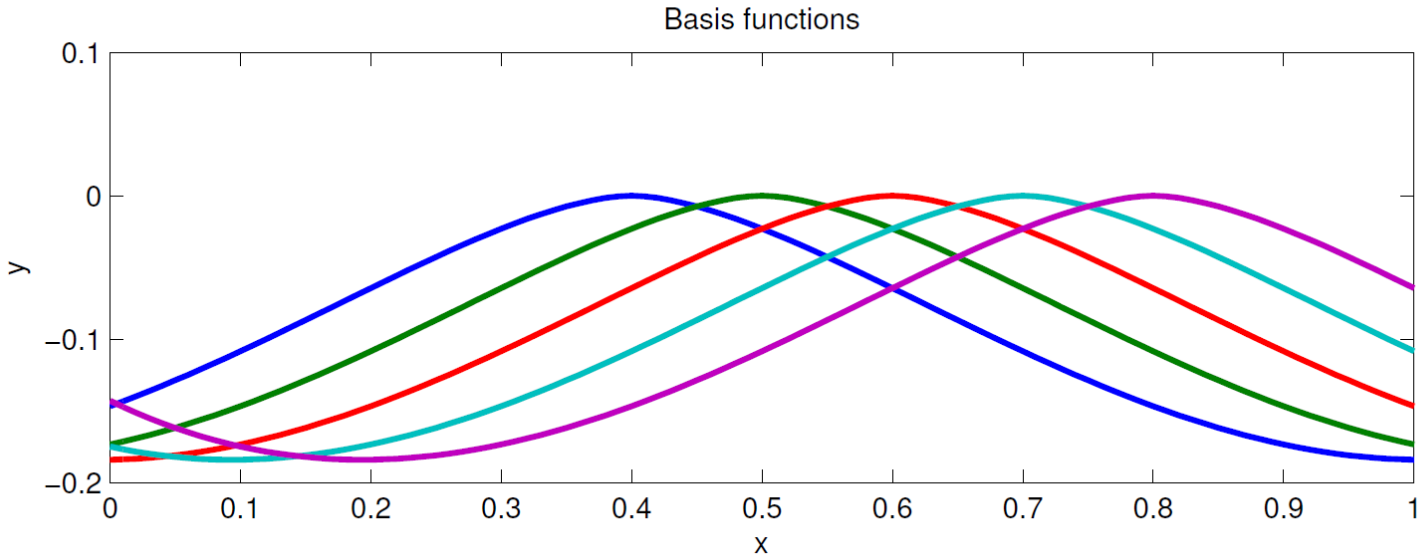
where

  - $(x - \xi_i)_+$ are called the basis functions,
  - ξ_i are called knots (the location of the basis function).

Flexible regression models with splines

↛ Spline example (single covariate with thinplate bases)





Flexible regression models

↛ Spline regression with multiple covariates

- Additive spline model
  - Each knot $\xi_{j.}$ (scaler) is connected with only one covariate

$$y = \alpha_0 + \alpha_1 x_1 + ... + \alpha_q x_q + \left[\sum_{j_1=1}^{m_1} \beta_{j_1} f(x_1, \xi_{j_1}) + ... + \sum_{j_q=1}^{m_q} \beta_{j_q} f(x_q, \xi_{j_q}) \right]$$

  - Good and simple if you know there is no interactions in the data a priori.
- Surface spline model
  - Each knot $\boldsymbol{\xi}_j$ (vector) is connected with more than one covariate

$$y = \alpha_0 + \alpha_1 x_1 + ... + \alpha_q x_q + \left[\sum_{j=1}^{m} \beta_j g(x_1, ...x_q, \boldsymbol{\xi}_j) \right] + \varepsilon$$

  - A popular choice of $g(x_1, ...x_q, \boldsymbol{\xi}_j)$ can be e.g. the multi-dimensional thinplate spline

$$g(x_1, ...x_q, \boldsymbol{\xi}_j) = \|\mathbf{x} - \boldsymbol{\xi}_j\|^2 \ln \|\mathbf{x} - \boldsymbol{\xi}_j\|$$

  - Can handle the interactions but the model complexity increase dramatically with the interactive knots

The challenges of using splines

- How many knots are needed?
  - Too few knots lead to a bad approximation; too many knots yield overfitting.
- Where to place those knots?
  - Equal spacing for the additive model,
  - which is obviously not efficient with the surface model.
- Common approaches to the two problems:
  - place enough many knots and use variable selection to pick up useful ones.
    - ⋆ not truly flexible
  - use reversible jump MCMC to move among the model spaces with different numbers of knots
    - ⋆ very sensitive to the prior and not computational efficient
  - clustering the covariates to select knots
    - ⋆ does not use the information from the responses
- How to choose between additive spline and surface spline?
  - NA

- The word "copula" means **linking**.
- **Sklar's theorem**
  Let H be a multi-dimensional distribution function with marginal distribution functions $F_1(x_1), ..., F_m(x_m)$. Then there exists a function C (**copula function**) such that

$$\begin{aligned} H(x_1, ..., x_m) =& C(F_1(x_1), ..., F_m(x_m)) \\ =& C\left(\int_{-\infty}^{x_1} f(z_1)dz_1, ..., \int_{-\infty}^{x_m} f(z_m)dz_m\right) = C(u_1, ..., u_m). \end{aligned}$$

  Furthermore, if $F_i(x_i)$ are continuous, then C is unique, and the derivative $c(u_1, ..., u_m) = \partial^m C(u_1, ..., u_m)/(\partial u_1 ... \partial u_m)$ is the **copula density**.

- If $X_1, ..., X_m$ are independent, and iff C is a product copula, then

$$C(F_1(x_1), ..., F_m(x_m)) = \prod_{i=1}^{m} F_i(x_i)$$

- The bivariate Gaussian copula

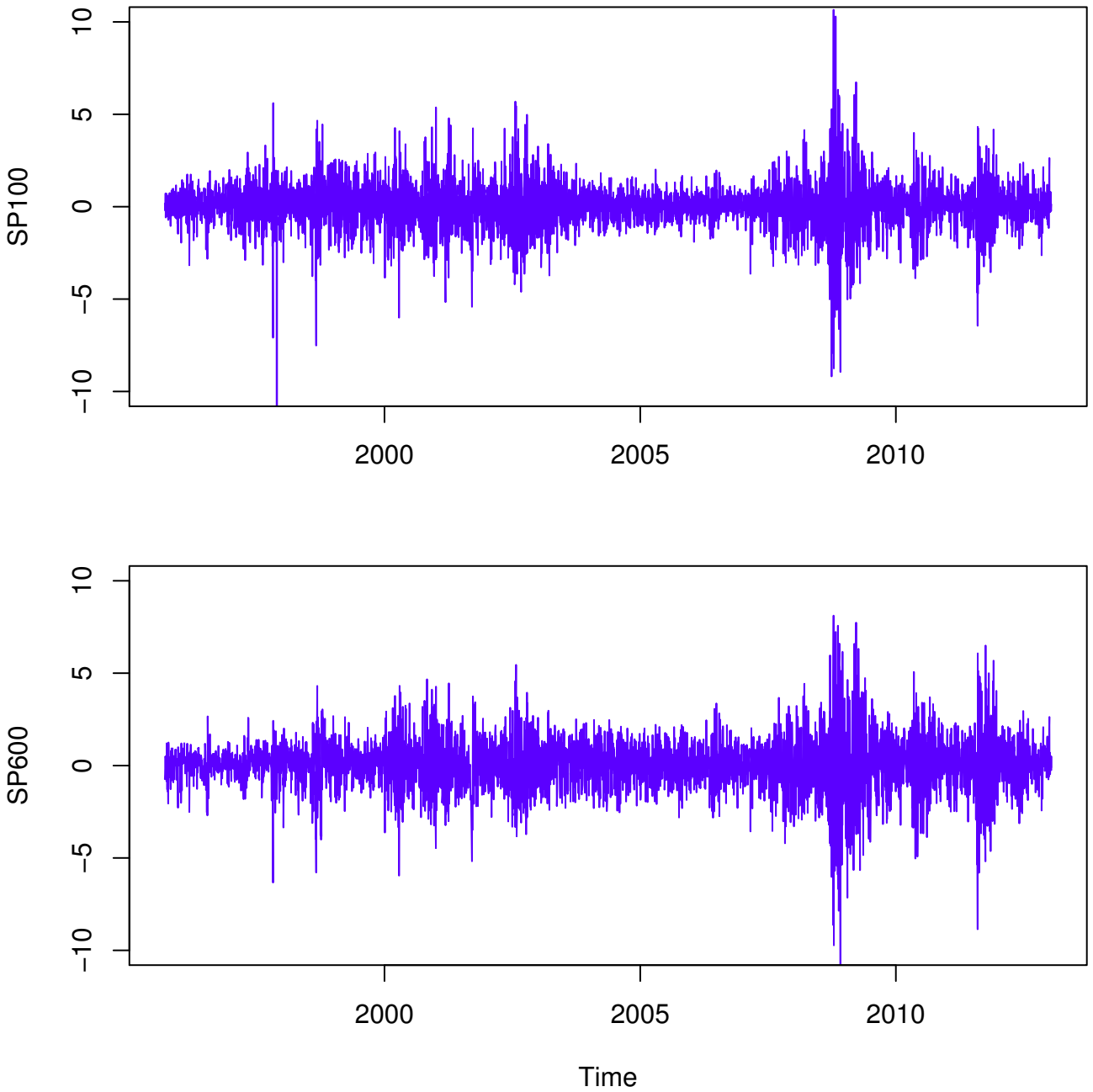
$$C(u_1, u_2, \rho) = \Phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \rho)$$
$$= \int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{z_1^2 - 2\rho z_1 z_2 + z_2^2}{2(1-\rho^2)}\right\} dz_1 dz_2$$

- The multivariate probit model is a simple example of a Gaussian copula, with univariate probit regressions as the marginals.
- There are many ways to construct a copula youself.

What can we do with the copula method?

- We can detect extreme events by observing some relative variables?
  - How many people are searching for the word "**flue**" in a certain area $\Rightarrow$ If there is a flue outbreak.
- We can construct a multivariate model that some margins are continuous but some are discrete?
  - One margin: A company's stock credited as A, A^+ over time.
  - The other margin: the stock returns over time
- In the *big data* world: we can estimate a very heavy multivariate model with the following steps
  - Independently build each marginal model. Parallel them!
  - Build the multivariate dependences on top of the margins.

The stock market returns



Measuring correlation and tail dependence

↛ Kendall's τ and tail-dependences

- The **Kendall's** τ can be written in terms of copula function:

$$\tau = 4\int\int F(x_1, x_2)\mathrm{d}F(x_1, x_2) - 1 = 4\int\int C(u_1, u_2)\mathrm{d}C(u_1, u_2) - 1.$$

- As well as the bivariate lower and upper **tail dependences**

$$\lambda_L = \lim_{u\to 0^+} \Pr(X_1 < F_1^{-1}(u)|X_2 < F_2^{-1}(u)) = \lim_{u\to 0^+} \frac{C(u,u)}{u},$$

$$\lambda_U = \lim_{u\to 1^-} \Pr(X_1 > F_1^{-1}(u)|X_2 > F_2^{-1}(u)) = \lim_{u\to 1^-} \frac{1-C(u,u)}{1-u}.$$

- Some facts:
  - The Kendall's τ is invariant w.r.t. **strictly** increasing transformations.
  - For all copulas in the elliptical class (Gaussian, t,...), $\tau = \frac{2}{\pi}\arcsin(\rho)$.
  - The Gaussian copula has zero tail dependence.
  - The student t copula has asmptotic upper tail dependence even for negative and zero correlations. The tail dependence decreases when degrees of freedom increases.

The covariate-contingent copula model

↛ The Joe-Clayton copula

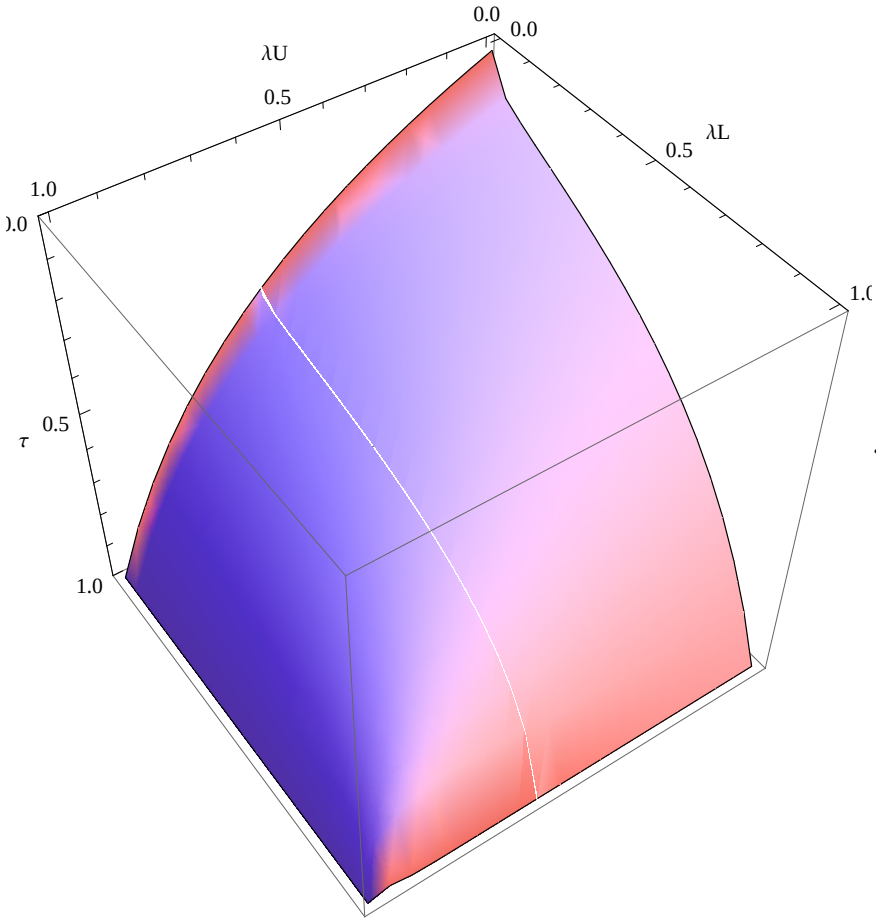
- The Joe-Clayton copula function

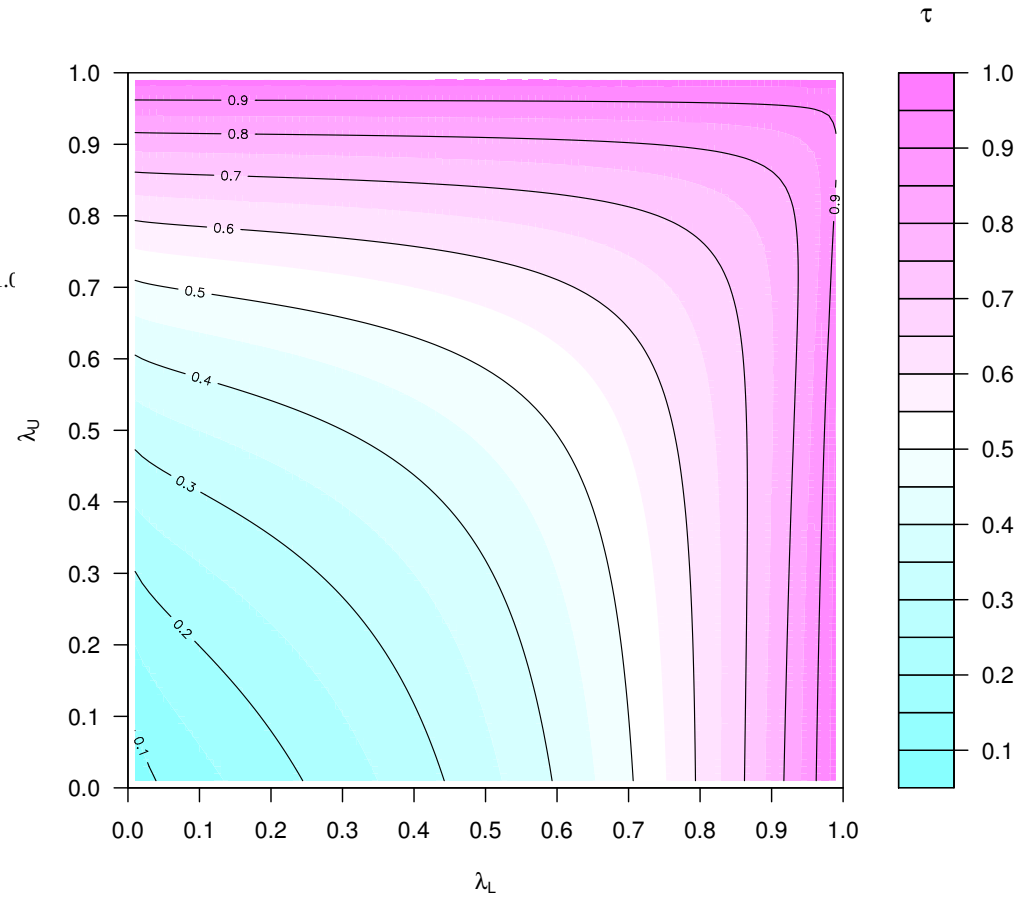
$$C(u, v, \theta, \delta) = 1 - \left[1 - \left\{\left(1 - \bar{u}^{\theta}\right)^{-\delta} + \left(1 - \bar{v}^{\theta}\right)^{-\delta} - 1\right\}^{-1/\delta}\right]^{1/\theta}$$

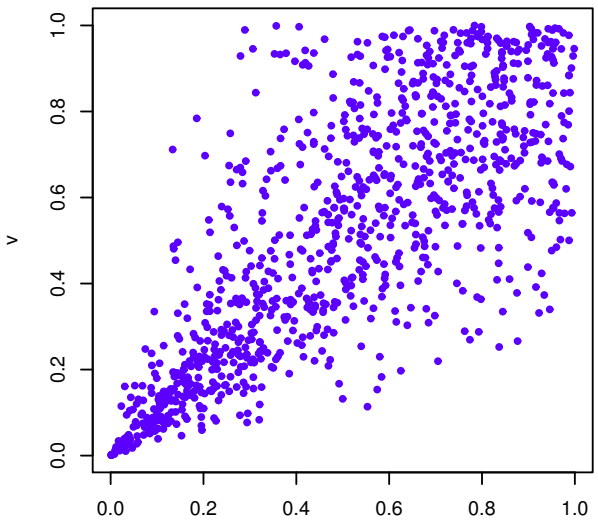
  where $\theta \geqslant 1$, $\delta > 0$, $\bar{u} = 1 - u$, $\bar{v} = 1 - v$.

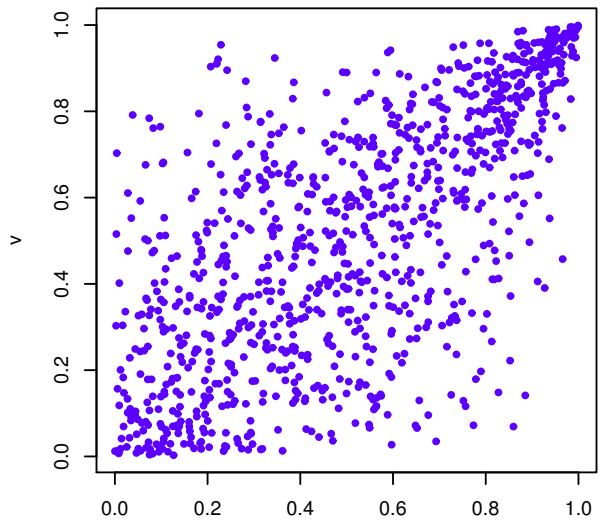
- Some properties:
  - One type of Archimedean copula.
  - $\lambda_L = 2^{-1/\delta}$ does not depend on $\lambda_U = 2 - 2^{-1/\theta}$.
  - $\tau = 1 - 4\int_0^{\infty} s \times (\varphi'(s))^2 ds$ is calculated via Laplace transform.
- Our interests:
  - The rank correlation and tail dependence in the model.
  - The convenience for interpretation.
- We use the reparameterized copula $C(u, v, \lambda_L, \tau) = C(u, v, \theta, \delta)$.

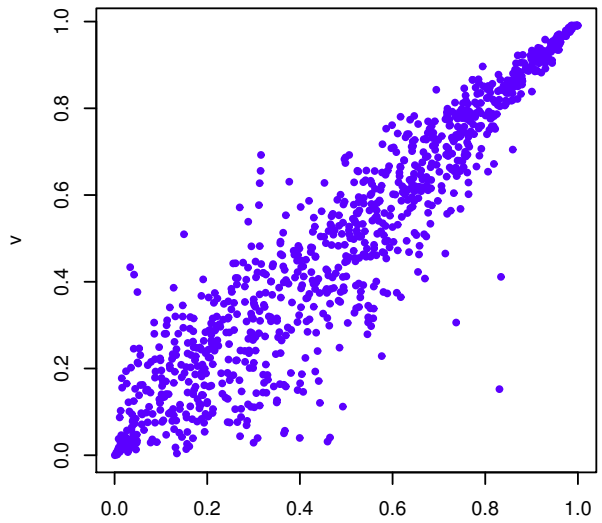
* **Note!** any other copulas can be equally well used.


λU
λL
τ
0.0
0.5
1.0


τ
λU
λL
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0


v
0.0
0.2
0.4
0.6
0.8
1.0


v
0.0
0.2
0.4
0.6
0.8
1.0


v
0.0
0.2
0.4
0.6
0.8
1.0

The covariate-contingent copula model

The model

- **The marginal models**
  - In principle, any combination of univariate marginal models can be used.
  - In the continuous case, we use univariate model that each margin is from the student t distribution.
- **The log likelihood**

$$\log \mathcal{L}(Y_u, Y_v | X_u, X_v, \lambda_L, \tau, \beta_u, \beta_v) = \sum_{i=1}^{n} \log c(u_i, v_i, \lambda_L, \tau) + \log \mathcal{L}_u(Y_u | X_u, \beta_u) + \log \mathcal{L}_v(Y_v | X_v, \beta_v)$$

where all the parameters are connected with covariates via link function $\varphi(\cdot)$, (identity, log, logit, probit,...)

The marginal features	$\beta_u = \varphi_{\beta_u}^{-1}(X_u \alpha_u)$	$\beta_v = \varphi_{\beta_v}^{-1}(X_v \alpha_v)$
The copula features	$\lambda_L = \varphi_{\lambda}^{-1}((X_u, X_v)\alpha_\lambda)$	$\tau = \varphi_{\tau}^{-1}((X_u, X_v)\alpha_$

The covariate-contingent copula model

↛ The fast approach

- Estimate each marginal model.
- Conditional on each marginal model, estimate the copula features