Penalized Splines -
A statistical Idea with numerous Applications ...

Göran Kauermann
Ludwig-Maximilians-University
Munich

Penalized Splines -

A statistical Idea with numerous Applications ...
which can be used for Copula Estimation

Göran Kauermann
Ludwig-Maximilians-University
Munich

Regression Splines in a Nutshell

The smooth regression model $\underline{Y = \mu(x) + \varepsilon}$ is fitted by replacing

$$\begin{aligned} \mu(x) &= \underbrace{\beta_0 + x\beta_1 + x^2\beta_2}_{} + \underbrace{\sum_{k=1}^{K} u_k (x - \tau_k)_+{}^2}_{} \\ &= \quad \boldsymbol{X}(x)\boldsymbol{\beta} \quad + \quad \boldsymbol{Z}(x)\boldsymbol{u} \quad = \quad \boldsymbol{B}(x)\boldsymbol{\theta} \end{aligned}$$

for knots $\tau_1, \tau_2, \ldots\ \tau_K$. This yields the estimate

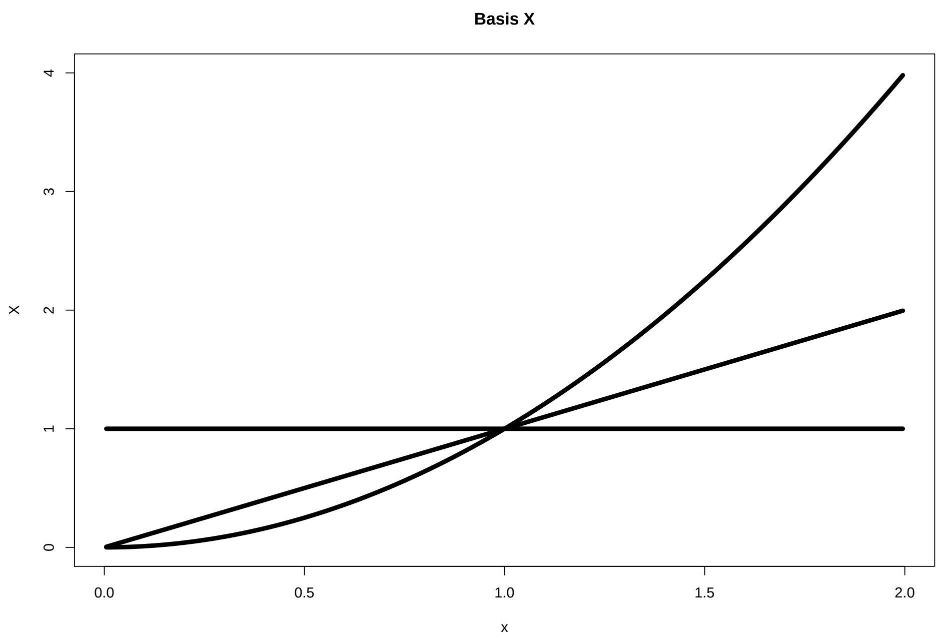
$$\widehat{\mu} = \boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{y}$$

$$\uparrow$$

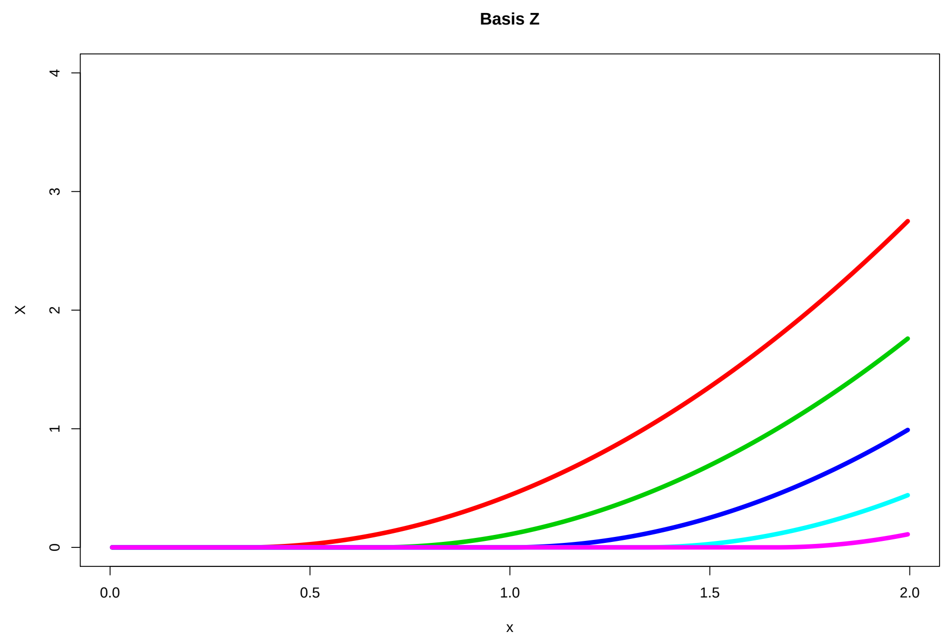
$(K+q) \times (K+q)$, with K "*large, but not too large*"

Regressions Splines

$X(x)$



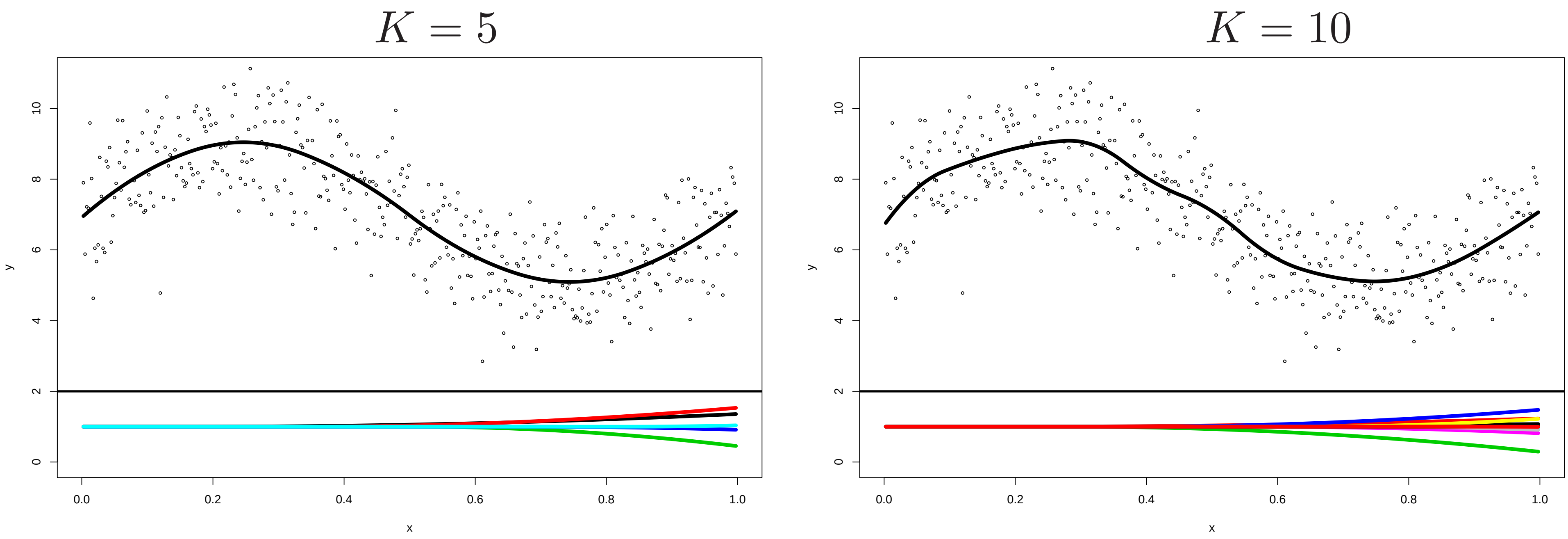
$Z(x)$



Regressions Splines

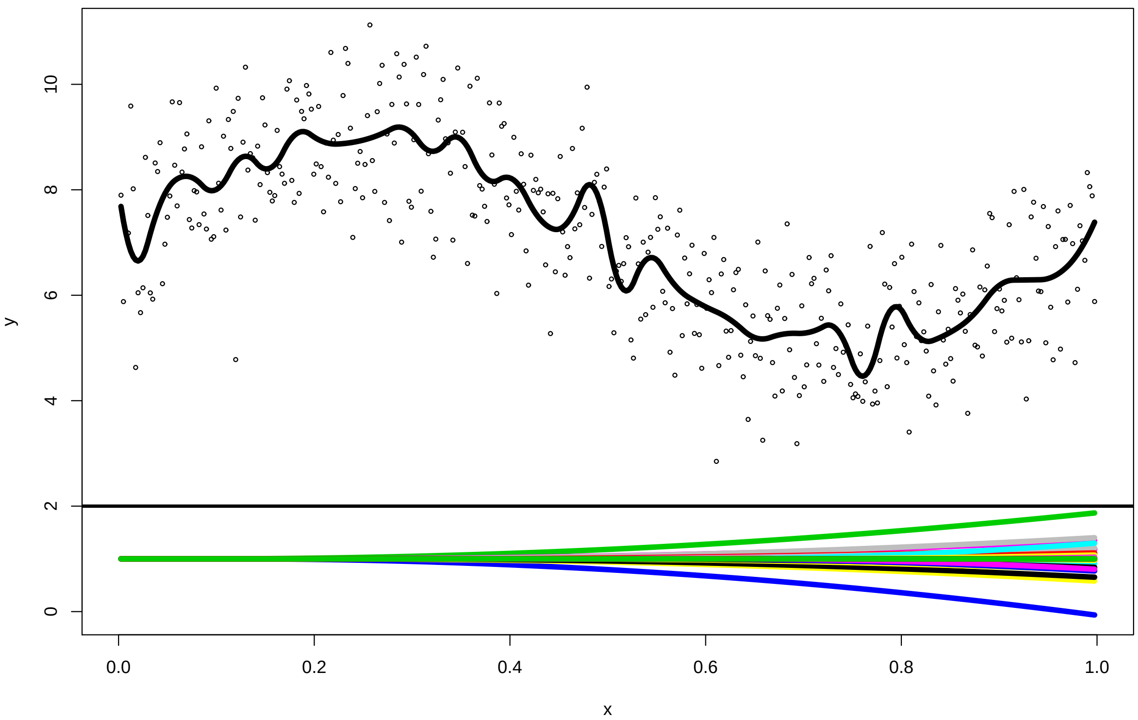
We get the estimate via

$$\widehat{\mu} \;=\; \boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{y}$$

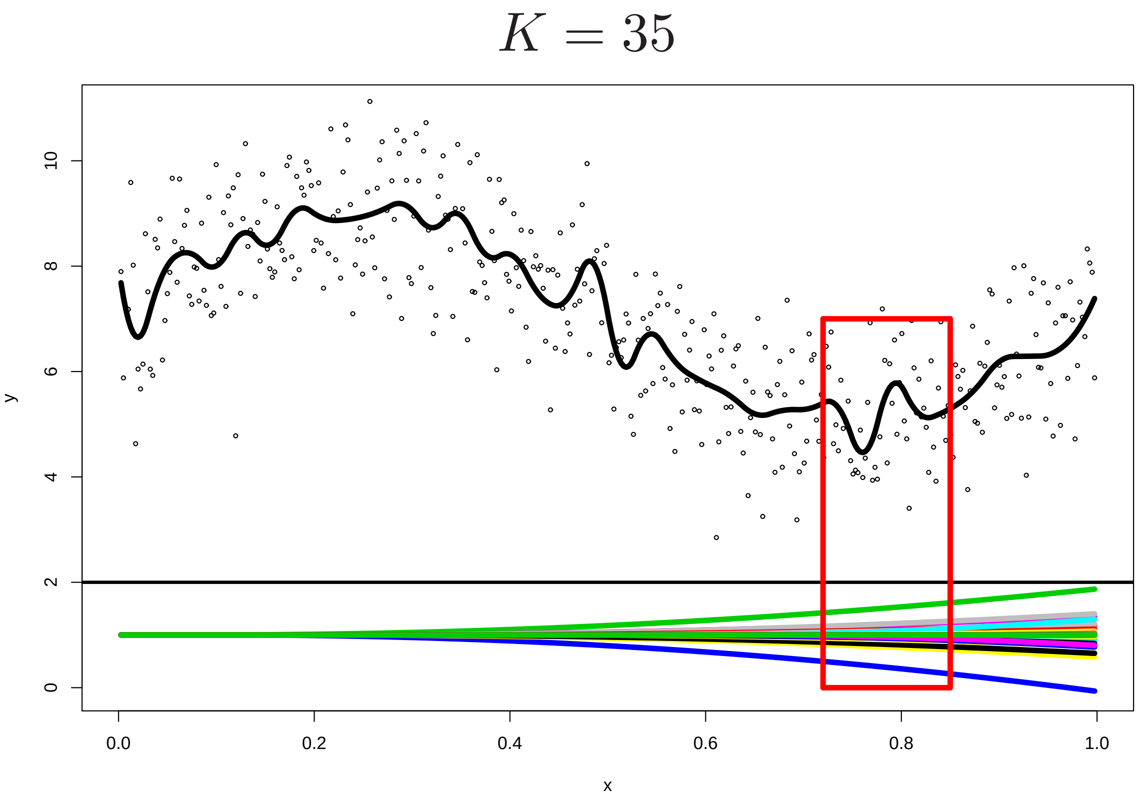


Need for Penalization

$K = 35$

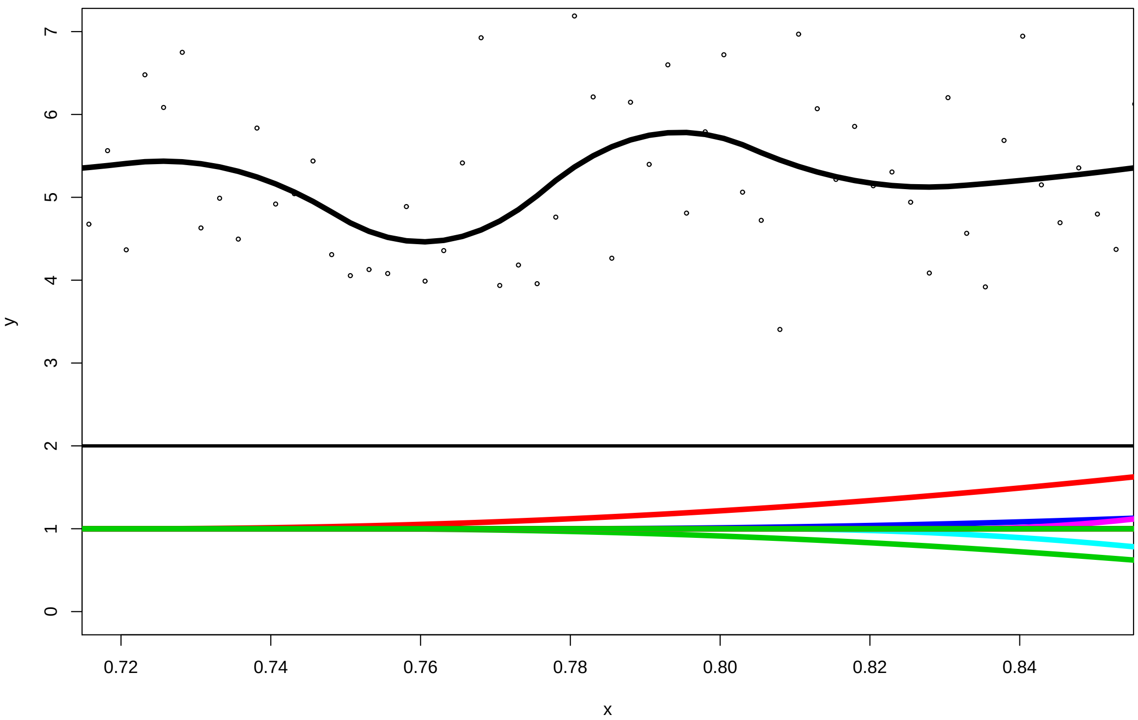


Need for Penalization



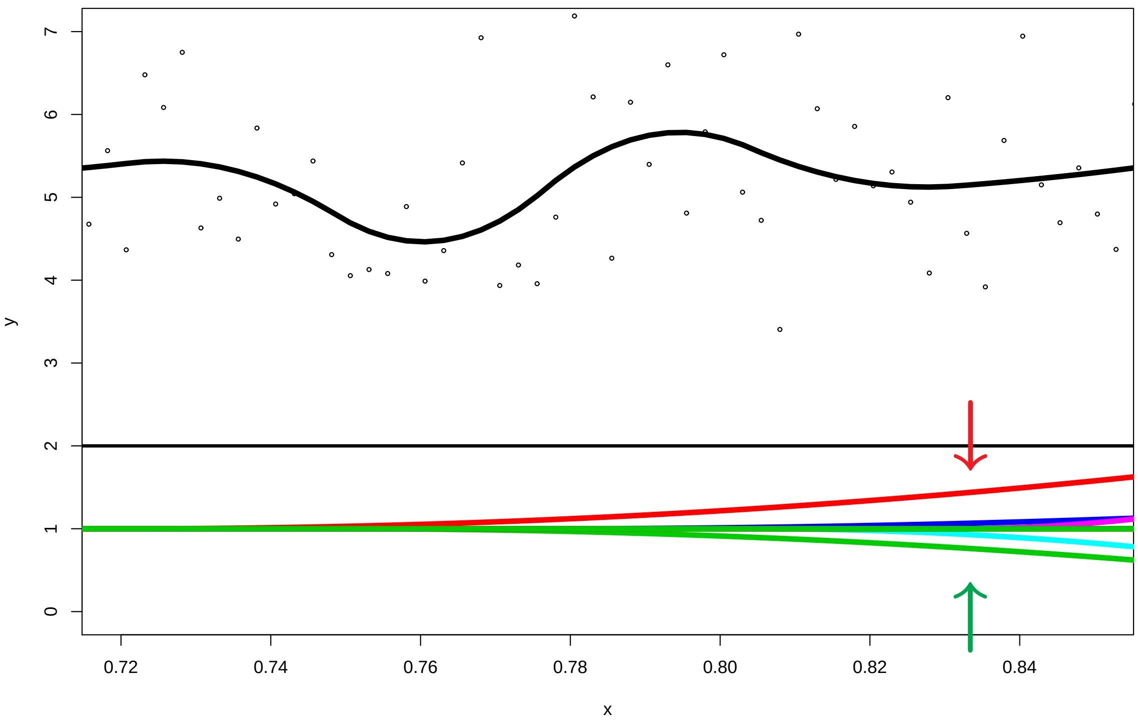
Need for Penalization

$K = 35$



Need for Penalization

$K = 35$



Penalized Least Square

We penalize the coefficients in $\boldsymbol{u}$ in that we postulate

$$\sum_{k=1}^{K} u_k^2 = \boldsymbol{u}^T \boldsymbol{I}_K \boldsymbol{u} \quad \longrightarrow \quad \text{small}$$

We minimize the Penalized Least Square

$$\sum_{i=1}^{n} \Big(y_i - \boldsymbol{X}(x_i)\boldsymbol{\beta} - \boldsymbol{Z}(x_i)\boldsymbol{u} \Big)^2 + \lambda \boldsymbol{u}^T \boldsymbol{I}_K \boldsymbol{u}$$

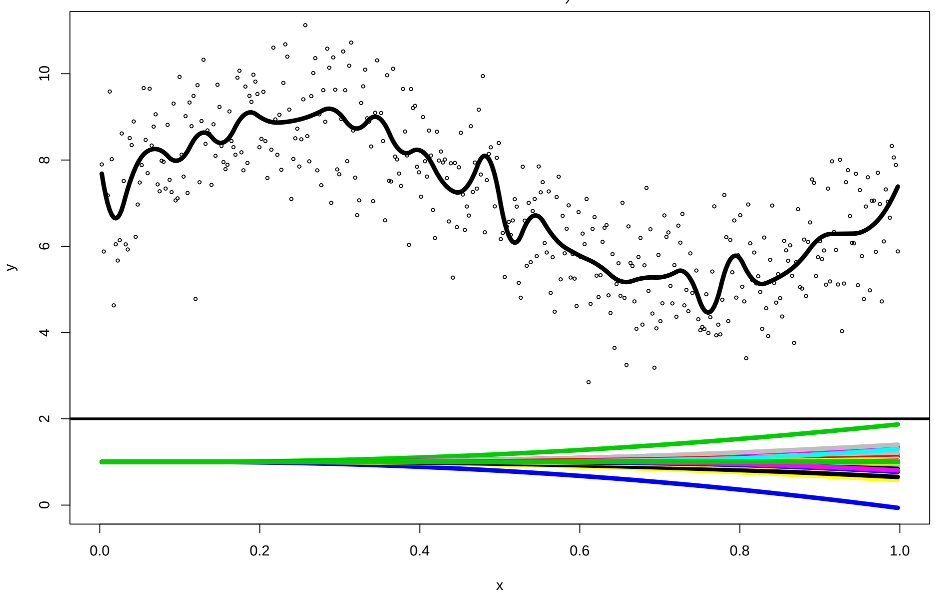
where λ is called the smoothing (penalization) parameter.

Penalized Regressions Splines

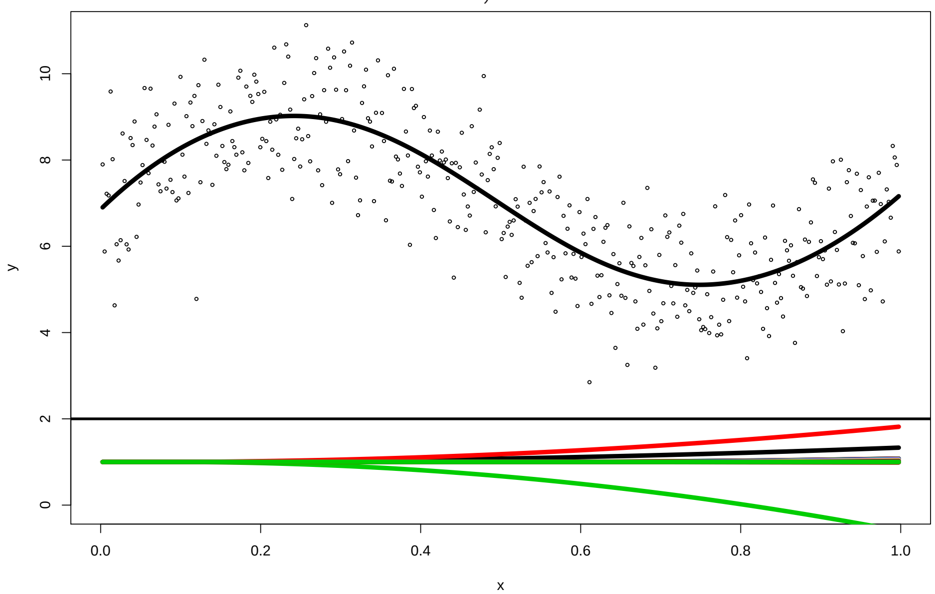
We get the estimate via

$$\widehat{\mu} \quad = \quad \boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B} + \boldsymbol{P}(\lambda))^{-1}\boldsymbol{B}^T\boldsymbol{y}$$

$K = 35, \lambda = 0$



$K = 35, \lambda = 1$



Penalized Spline Recipe

(O'Sullivan, 1986, Eilers & Marx, 1996, Ruppert, Wand & Carroll 2003)

"The Penalized Spline Recipe":

1. Take a **rich, high dimensional** Basis $\boldsymbol{B}(x)$, i.e. choose K generously large.

2. Minimize the penalized least squares criterion

$$(\boldsymbol{Y} - \boldsymbol{B\theta})^T(\boldsymbol{Y} - \boldsymbol{B\theta}) + \lambda \boldsymbol{\theta}^T \boldsymbol{D\theta} \quad \rightarrow \quad \text{min}$$

   with $\boldsymbol{D}$ as adequately chosen **penalty matrix**, in the simplest case $\boldsymbol{D} = \text{diag}(\boldsymbol{0}_q, \boldsymbol{I}_K)$

3. Choose penalty parameter λ **data-driven** (e.g. by cross validation)

Reformulation

For general splines can rewrite the penalized estimation to
(Wand & Ormerod, 2008, Aust. & NZ J. Stat)

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}$$

with $\boldsymbol{X}$ low dimensional and $\boldsymbol{Z}$ high dimensional and penalized least square

$$r(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u}) + \lambda \boldsymbol{u}^T\boldsymbol{D}\boldsymbol{u} \quad \rightarrow \quad \min_{u,\beta}$$

with penalty matrix $\boldsymbol{D}$ usually chosen as identity matrix.

Reformulation

For general splines can rewrite the penalized estimation to
(Wand & Ormerod, 2008, Aust. & NZ J. Stat)

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\epsilon}$$

with $\boldsymbol{X}$ low dimensional and $\boldsymbol{Z}$ high dimensional and penalized least square

$$(\boldsymbol{Y} - \boldsymbol{X\beta} - \boldsymbol{Zu})^T(\boldsymbol{Y} - \boldsymbol{X\beta} - \boldsymbol{Zu}) + \lambda \underbrace{\boldsymbol{u}^T\boldsymbol{Du}}_{\text{quadratic form}} \quad \rightarrow \quad \min_{u,\beta}$$

with penalty matrix $\boldsymbol{D}$ usually chosen as identity matrix.

Linking Penalized Splines with Linear Mixed Models

We formulate the penalty as *a priori* normal distribution:

$$\boldsymbol{u} \quad \sim \quad N(\mathbf{0}, \sigma_u^2 \boldsymbol{D}^{-1}) \qquad (1)$$

Now, coefficient vector $\boldsymbol{u}$ is considered as random.

Conditioning on $\boldsymbol{u}$ yields

$$\boldsymbol{Y}|\boldsymbol{u} \quad \sim \quad N(\boldsymbol{X\beta} + \boldsymbol{Zu}, \sigma_\epsilon^2 \boldsymbol{I}) \qquad (2)$$

With (1) and (2) we get a **Linear Mixed Model**

Posterior Estimates in a Linear Mixed Models

We ”estimate” $\boldsymbol{u}$ through the **Posterior Bayes estimate**

(or equivalently Best Linear Unbiased Predictor (BLUP))

$$\widehat{\boldsymbol{u}} \;=\; E(\mathbf{u}|\boldsymbol{Y};\boldsymbol{\beta}) \;=\; \left(\boldsymbol{Z}^T\boldsymbol{Z} + \frac{\sigma_\varepsilon^2}{\sigma_u^2}\boldsymbol{D}\right)^{-1}\boldsymbol{Z}^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$$

where the penalty parameter equals

$$\lambda = \sigma_\varepsilon^2/\sigma_u^2$$

Penalized Spline Smoothing and Linear Mixed Models

We obtain the following results:

- This Posterior Bayes estimate is equivalent to the Penalized Spline estimate

- The penalty parameter $\lambda = \sigma_\varepsilon^2 / \sigma_u^2$ is a regular parameter in the linear mixed model.

- Smoothing (with penalized splines) can be carried out with software for fitting linear mixed models

Collection of Results

- Smoothing parameter $\lambda = \sigma_\epsilon^2/\sigma_u^2$ can be estimated by maximum likelihood.

  $\Rightarrow$ This avoids grid searching!

  (Kauermann, 2005, *JSPI*; Ruppert, Wand & Carroll, 2003).

- Smoothing parameter can be selected in the presence of correlated errors

  $\Rightarrow$ AIC based selection fails here!

  (Krivobokova & Kauermann, 2007, *JASA*)

Collection of Results (cont)

- Asymptotic results on the number of knots

  $\Rightarrow$ Penalized Splines are asymptotically justified!

  (Kauermann, Krivobokova & Fahrmeir, 2009, *JRSS B* ...)

- Number of knots for fixed sample size

  $\Rightarrow$ Practical decision rule on how to choose K

  (Kauermann & Opsomer 2011, *Biometrika*)

Collection of Results (cont)

- Local adaptive smoothing

  $\Rightarrow$ Simple and fast computation!

  (Krivobokova, Crainiceanu & Kauermann, 2008, *JCGS*)

- .Small area estimation and smoothing

  $\Rightarrow$ Combination of smoothing and mixed models!

  (Opsomer, Claeskens, Ranalli, Kauermann & Breidt, 2008, *JRSS B*)

- and ...

Outline of (Rest of) Talk

Penalized Spline Fitting is well applicable beyond regression!!

We show how to use Penalized Splines for **Copula Estimation**

- The idea of Sparse Grids
- The idea of Pair-Copulas

Penalized Smooth Estimation of Copulas

using Sparse Grids

Copulas

The idea of copulas traces back to: Hoeffding (1940), Sklar (1959)

In 1997 first entry for ’copula’ in *Encyclopedia of Statistical Sciences*

Wide area of applications and theory:

- mathematics, e.g. Joe (1997) or Nelsen (2006)
- financial econometrics, e.g. Embrechts (2009)
- biostatistics, e.g. Bogaerts & Lesaffre (2008)
- marketing, e.g. Danaher & Smith (2011)
- engineering, e.g. Kelly (2007)
- ecology, e.g. Briggs et al. (2011)
- . . .

Definition of Copulas

Sklar's (1959) theorem: The distribution function of a p dimensional random vector $x = (x_1, \ldots, x_p)$ can be written as

$$F(x_1, \ldots, x_p) = C\{F_1(x_1), \ldots, F_p(x_p)\},$$

We assume that the $C(\cdot)$ is differentiable, so that the density results as

$$f(x_1, \ldots, x_p) = c(F_1(x_1), \ldots, F_p(x_p)) \prod_{j=1}^{p} f_j(x_j).$$

The copula carries the multivariate dependence structure

The idea of Copulas

A multivariate distribution $F(x_1, \ldots, x_p)$ decomposes into:

1. The p univariate marginal distributions $F_1(x_1), \ldots F_p(x_p)$
2. The dependence structure in form of the copula density

$$c(u_1, \ldots, u_p)$$

   on the unit cube, i.e. $u_j \in [0, 1]$, $j = 1, \ldots, p$.

Our task: Estimation of $c(\cdot)$ with penalized splines.

Properties of the Copula Density

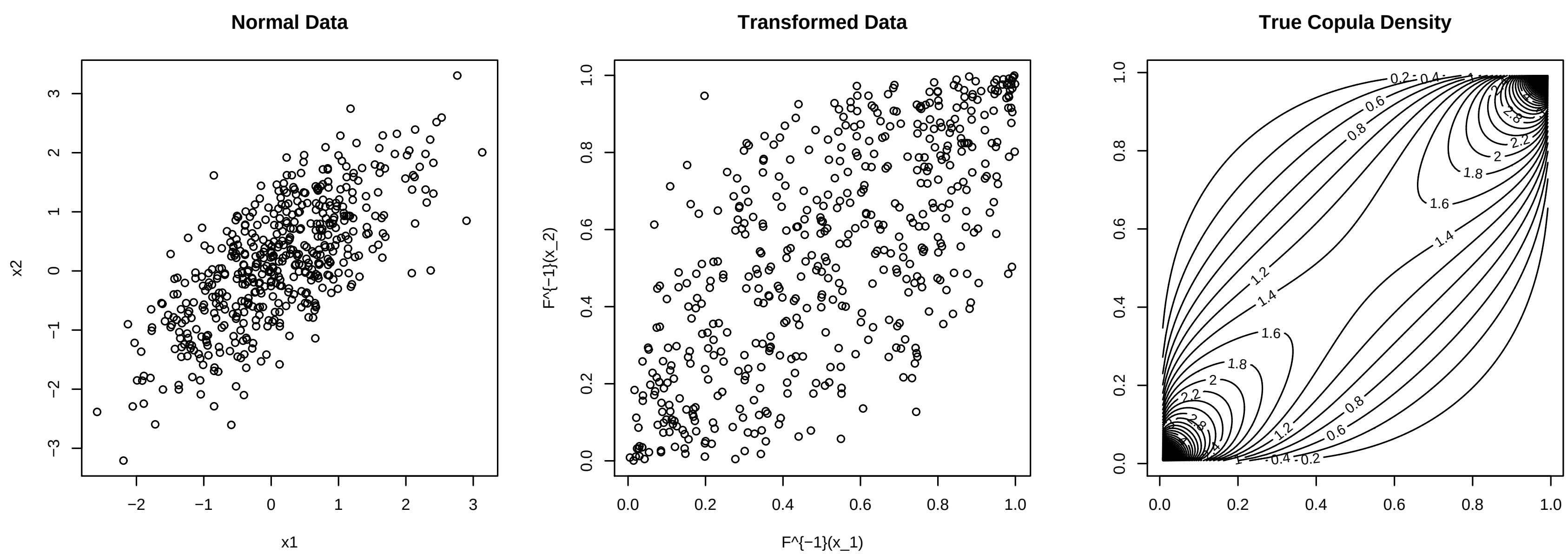
- The copula density $c(u_1, \ldots, u_p)$ has the bounded support $[0, 1]^p$.

- Univariate margins of $c(u_1, \ldots, u_p)$ are uniform, i.e.

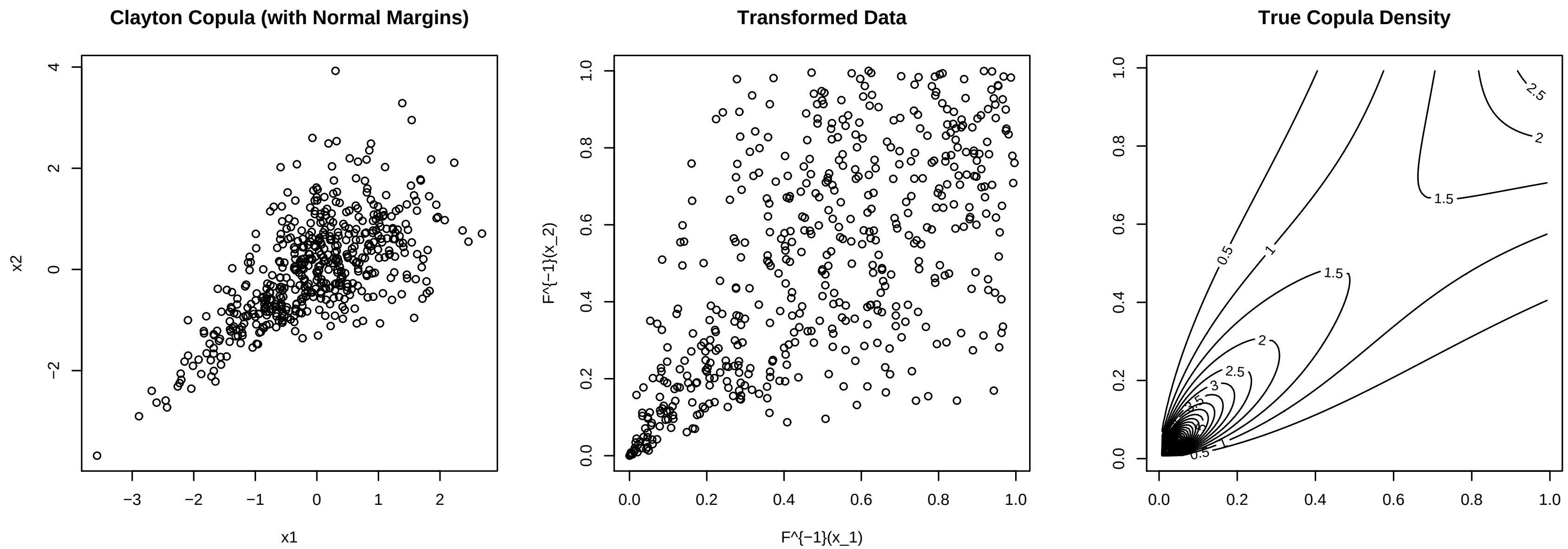
$$c_j(u_j) \quad \equiv \quad 1 \text{ for } j = 1, \ldots p$$

- Often, high density areas are at the boundary of $[0, 1]^p$.

Demonstration - Normal Copula



Demonstration - Clayton Copula



Nonparametric Copula Estimation

Nonparametric copula estimation is weakly developed (probably) since

1. Constraints: The copula density has uniform margins, this is hard to accommodate in classical kernel density estimation.

2. Boundary: The support is bounded, which requires special kernels to avoid boundary bias problems.

3. Dimension: The copula idea is suitable for high dimensions, nonparametric density estimation is not suited for that (*curse of dimensionality*).

Penalized Estimation of a Copula

Penalized Estimation of Copulas tackles the three problems:

1. Constraints: We will use B-splines and Bernstein polynomials, which easily accommodate the constraints.

2. Boundary: The splines are bounded, so there is no boundary problem.

3. Dimension: We tackle the curse of dimensionality with ’sparse grids’ and ’pair-copulas’.

B-spline fitting of Copulas

Let $u_j = F_j^{-1}(x_j)$ so that $c(u_1, \ldots, u_p)$ is a density on $[0,1]^p$.

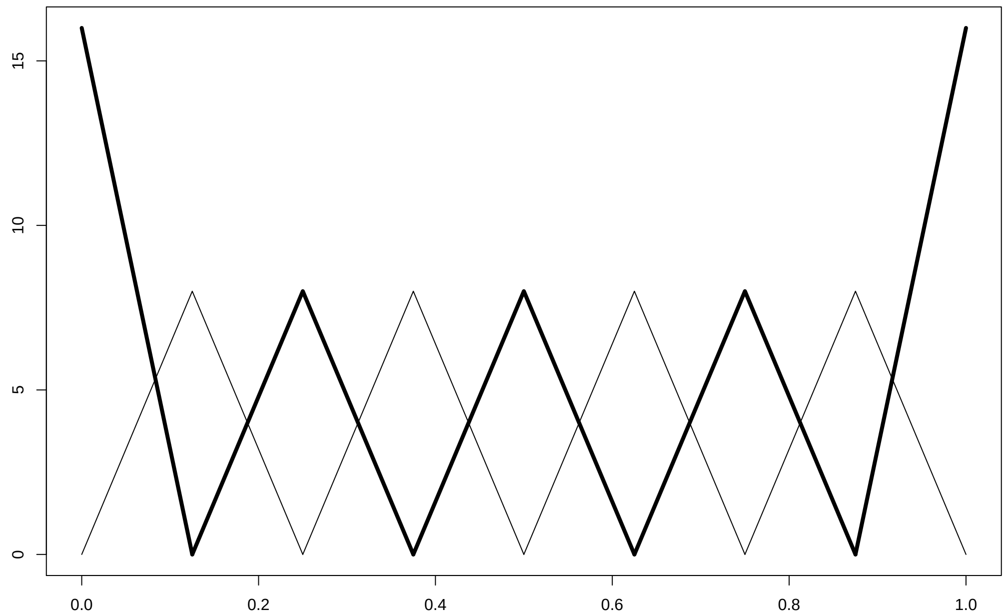
Let $k = (k_1, \ldots, k_p) \in \mathcal{K}$ be a p-dimensional multi index.

We replace/approximate $c(\cdots)$ by

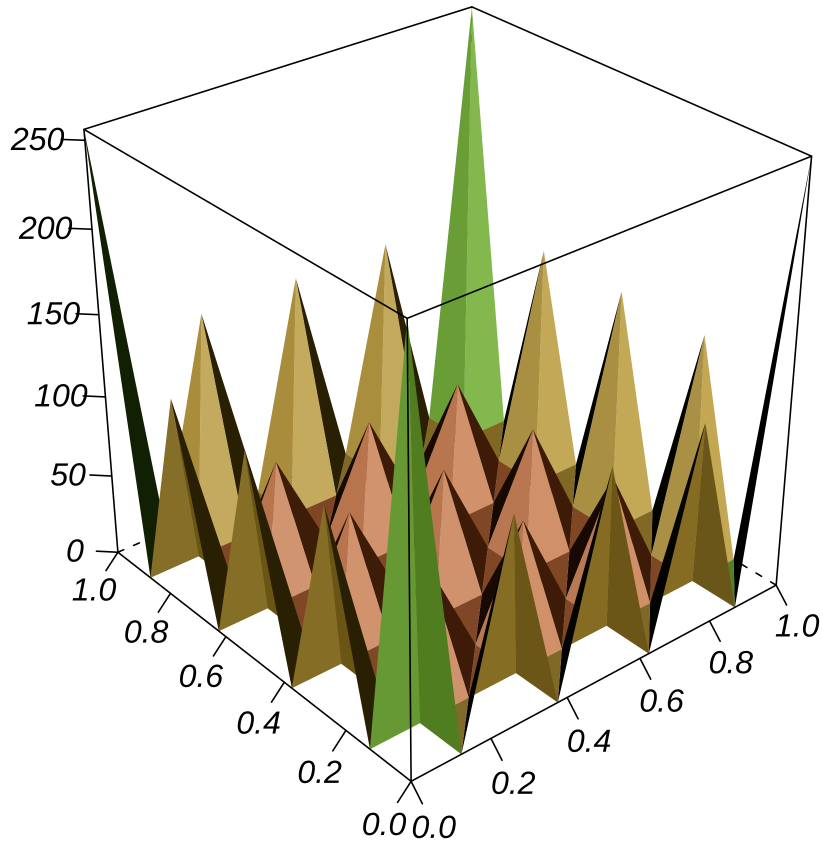
$$\boxed{c(u_1, \ldots, u_p) \approx \sum_{k \in \mathcal{K}} \boldsymbol{\phi}(u_1, \ldots, u_p)\, b_k =: \sum_{k \in \mathcal{K}} \prod_{j=1}^{p} \phi_{k_j}(u_j) b_k,}$$

where $\phi_l(\cdot)$ are univariate B-spline Density Bases

Marginal B-spline Density Basis



Tensor Product Basis



Log Likelihood

We assume i.i.d. data $u_i = (u_{i1}, \ldots, u_{ip})$, for $i = 1, \ldots, n$
The log likelihood equals

$$l(\boldsymbol{b}) = \sum_{i=1}^{n} \log \left(\sum_{k \in \mathcal{K}} \phi_k(u_i) b_k \right)$$

with $\boldsymbol{b} = (b_k, k \in \mathcal{K})$.

We approximate $l(\boldsymbol{b})$ by second order Taylor series, i.e.

$$l(\boldsymbol{b}) = \underbrace{s(\boldsymbol{b}) + \boldsymbol{b}^T \boldsymbol{H} \boldsymbol{b}}_{=: \underline{Q(\boldsymbol{b})}} + \ldots$$

and we maximize $Q(\boldsymbol{b})$.

Penalizing the log Likelihood

We need a penalization to obtain a smooth fit.

Penalized likelihood:

$$
\begin{aligned}
l_p(\boldsymbol{b}, \lambda) &= l(\boldsymbol{b}) - \underbrace{\frac{1}{2}\boldsymbol{b}^T \boldsymbol{D}(\lambda)\boldsymbol{b}}_{\text{Penalty}} \\
&\approx Q(\boldsymbol{b}) - \frac{1}{2}\boldsymbol{b}^T \boldsymbol{D}(\lambda)\boldsymbol{b} =: Q_p(\boldsymbol{b}, \lambda)
\end{aligned}
$$

with penalty matrix $\boldsymbol{D}(\lambda) = \sum_{j=1}^{p} \lambda_j D_j$

Constraints on the Parameters

1. The marginal density is **uniform** which results with the linear constraint

$$c_j(u_j) = \sum_{k_j}^{K} \phi_{k_j}(u_j) b_{(j)k_j} = 1 \tag{3}$$

   with $b_{(j)k_j}$ as marginal spline coefficient.

2. We fit a density with $\int c(u)\mathrm{d}u = 1$, which results through

$$\sum_{k \in \mathcal{K}} b_k = 1 \tag{4}$$

3. The density is positive, which results with

$$\sum_{k \in \mathcal{K}} \phi_k(u_1, \ldots, u_p) b_k \geq 0 \tag{5}$$

Quadratic Programming

We intend to maximize

$$Q_p(\boldsymbol{b}, \lambda) \to \text{max}$$

subject to the given linear constraints (previous slide), written as

$$\boldsymbol{A}\boldsymbol{b} = \mathbf{1}, \quad \boldsymbol{B}\boldsymbol{b} \geq \mathbf{0}$$

This can be solved with (iterative) Quadratic Programming
(R package `quadprog`).

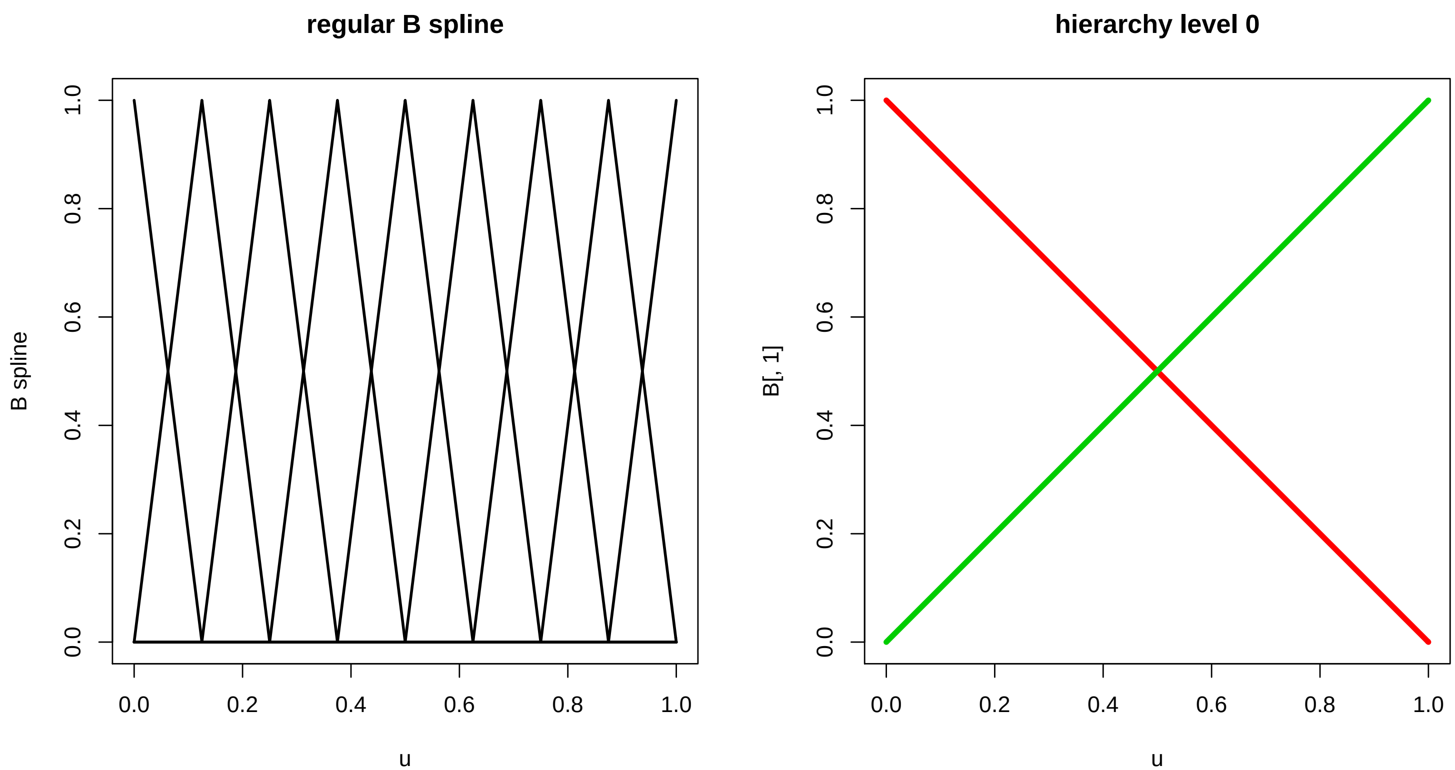
The Curse of Dimensionality

For $p > 2$ the dimension of the full tensor product basis becomes numerically infeasible.

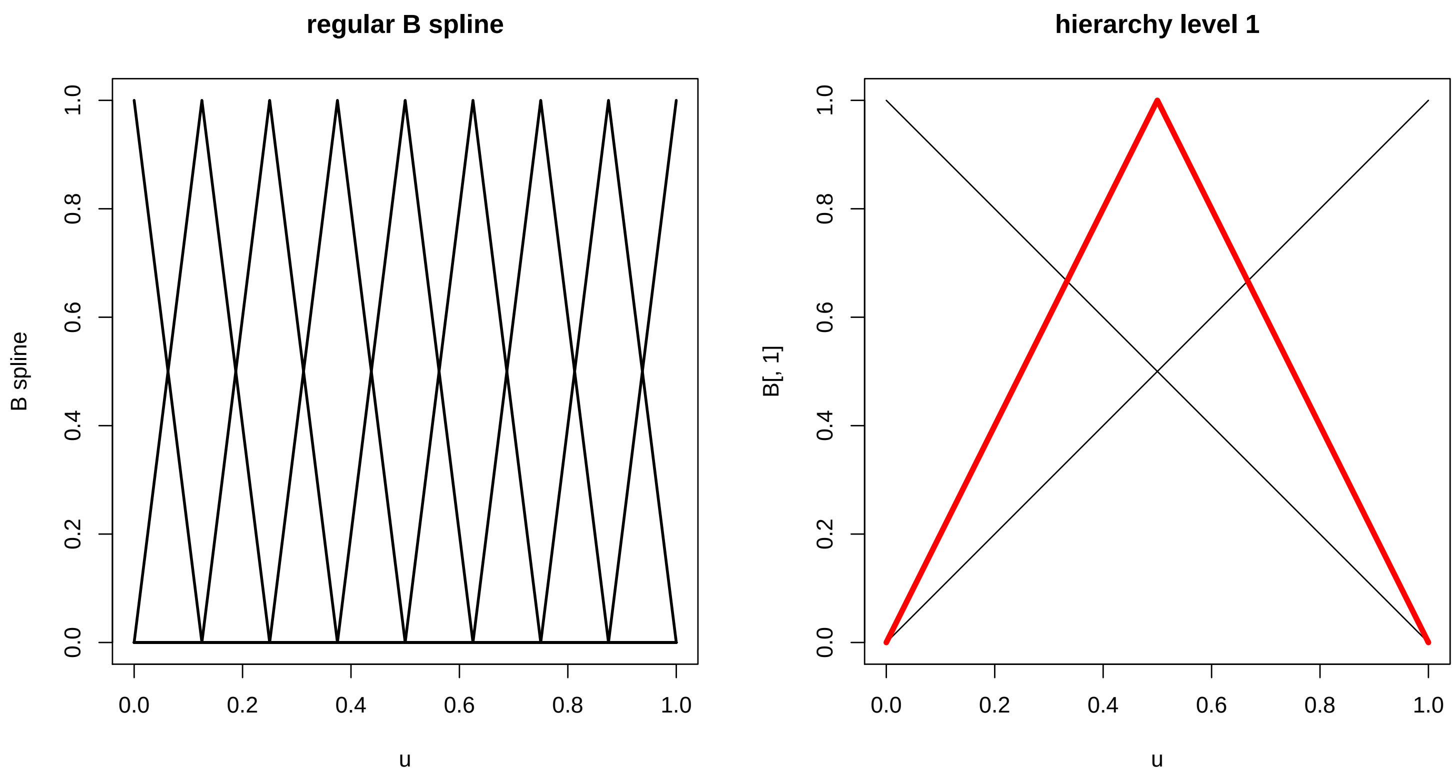
Dimension of Spline Basis

marginal; spline dimension	basis	$p = 3$	$p = 4$	$p = 5$
K=9 (2^3)	tensor product	**729**	**6,561**	**59,049**
K=17 (2^3)	tensor product	**4,913**	**83,521**	**1,419,857**
K=33 (2^3)	tensor product	**35,937**	**1,185,921**	**39,135,393**

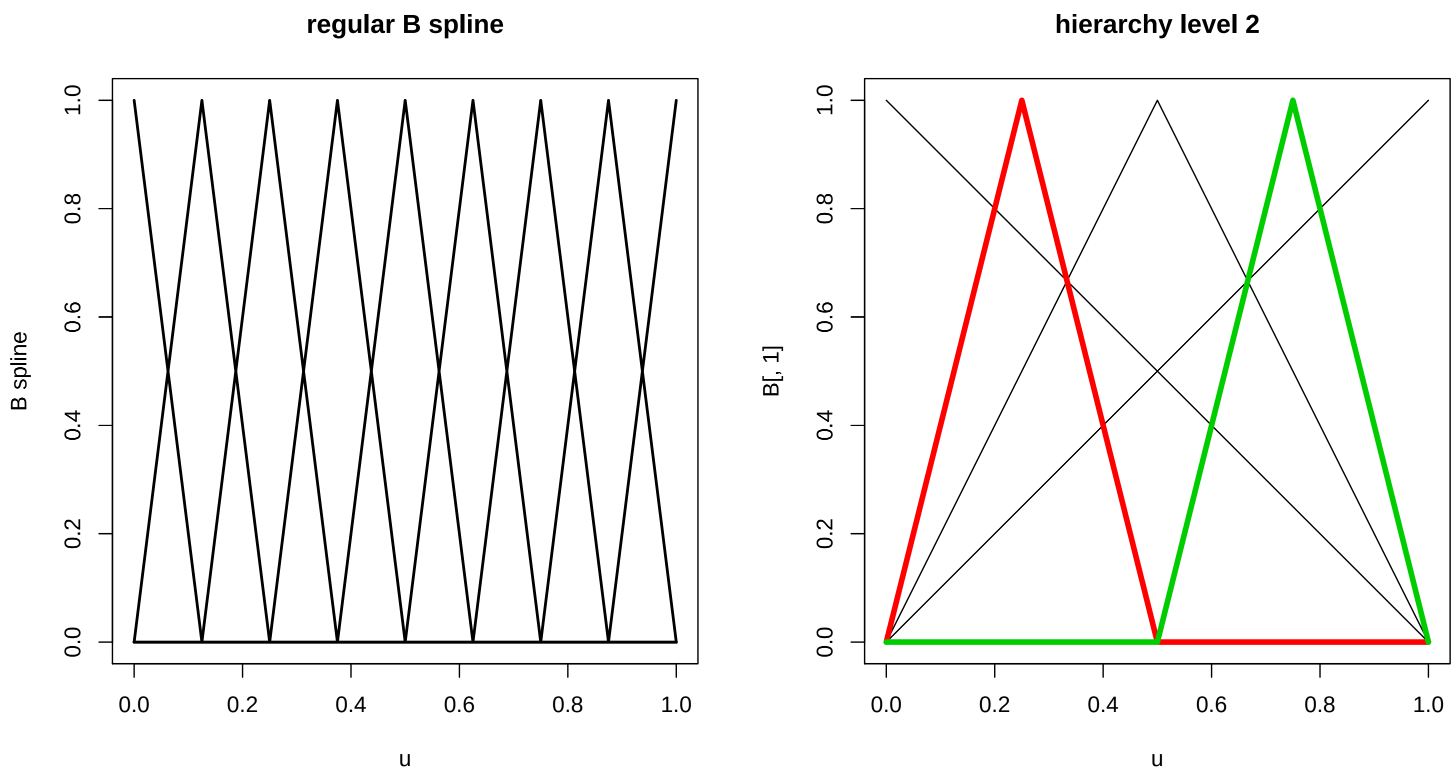
Hierarchical B-splines



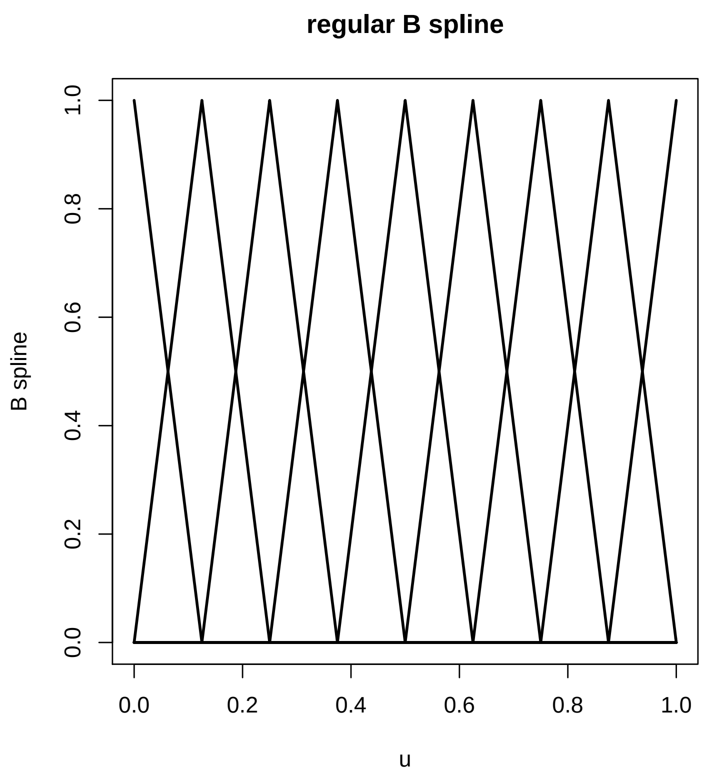
Hierarchical B-splines

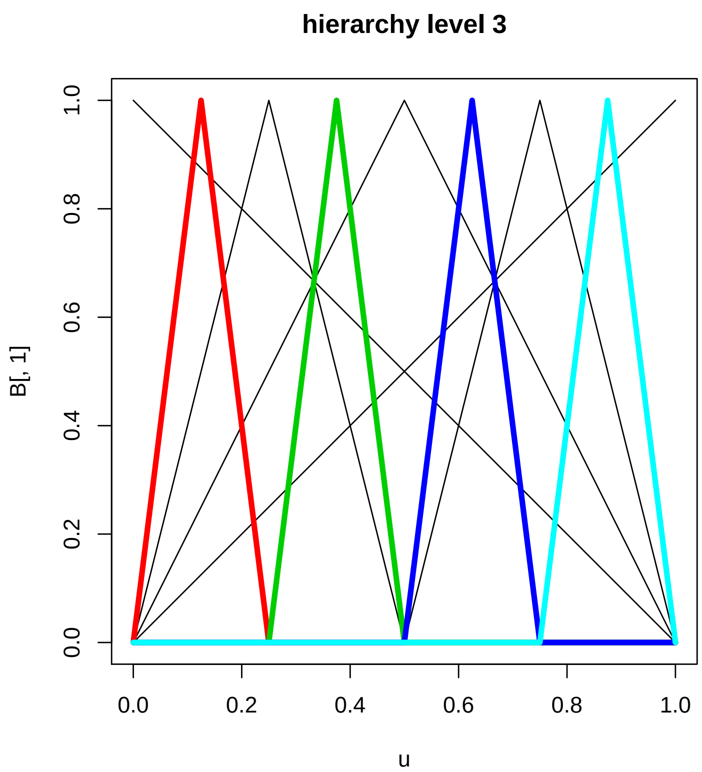


Hierarchical B-splines



Hierarchical B-splines

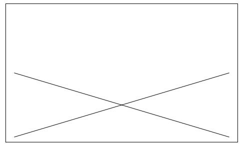
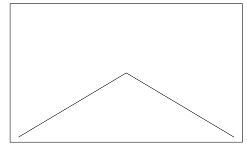
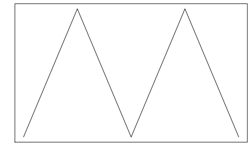
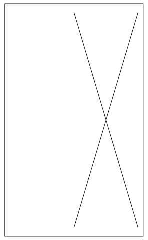
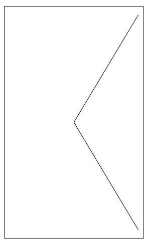
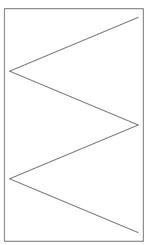




Tackling the Curse of Dimensionality

The idea is now to built a sparse tensor product, by including tensors up to a given hierarchy only.

Sparse grids have been proposed by Zenger (1992, Notes on Numerical Fluid Mechanics)

			
	$\Phi_{(0)}(u_1) \otimes \Phi_{(0)}(u_2)$	$\Phi_{(1)}(u_1) \otimes \Phi_{(0)}(u_2)$	$\Phi_{(2)}(u_1) \otimes \Phi_{(0)}(u_2)$
	$\Phi_{(0)}(u_1) \otimes \Phi_{(1)}(u_2)$	$\Phi_{(1)}(u_1) \otimes \Phi_{(1)}(u_2)$	neglected
	$\Phi_{(0)}(u_1) \otimes \Phi_{(2)}(u_2)$	neglected	neglected

Tackling the Curse of Dimensionality

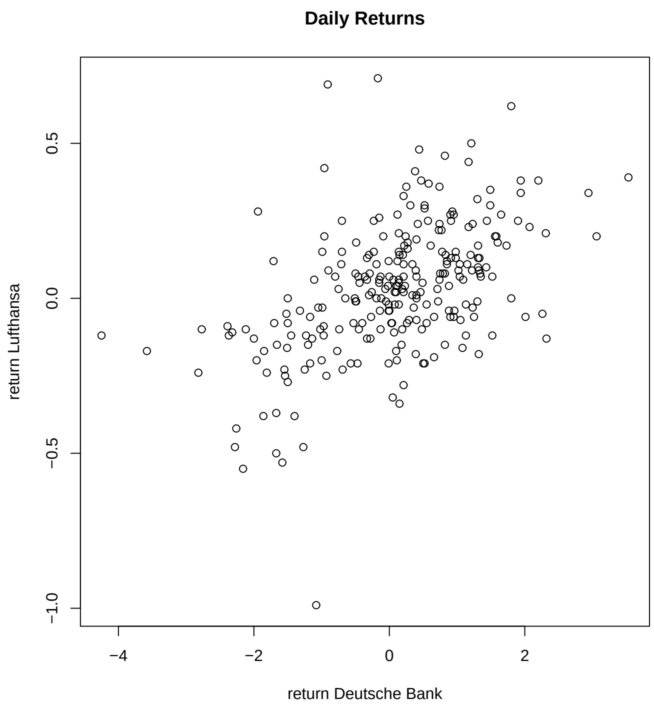
Sparse grids allow to push the Curse of Dimensionality

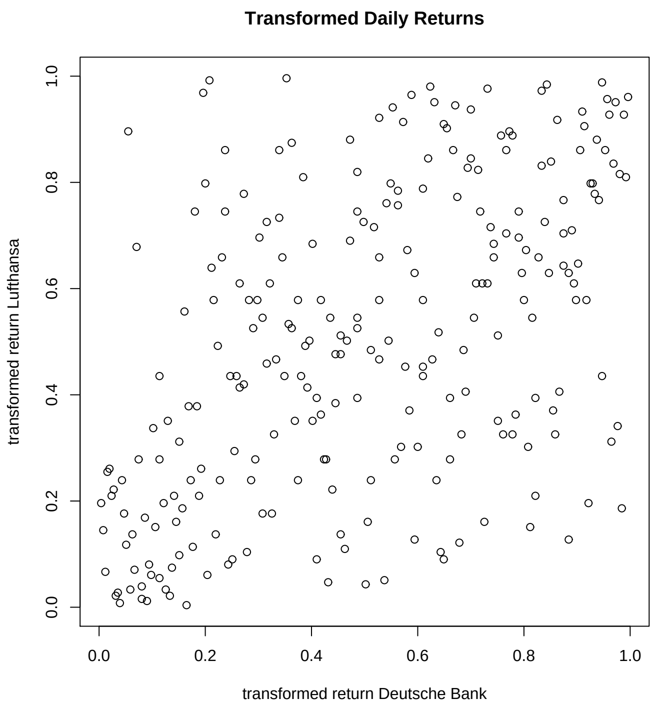
Dimension of Spline Basis

margin; spline dimension	basis	$p = 3$	$p = 4$	$p = 5$
K=9 (2^3)	tensor product	729	6,561	59,049
	sparse grid basis	**123**	**297**	**705**
K=17 (2^3)	tensor product	4,913	83,521	1,419,857
	sparse grid basis	**368**	**961**	**2,441**
K=33 (2^3)	tensor product	35,937	1,185,921	39,135,393
	sparse grid basis	**1,032**	**2,882**	**7,763**

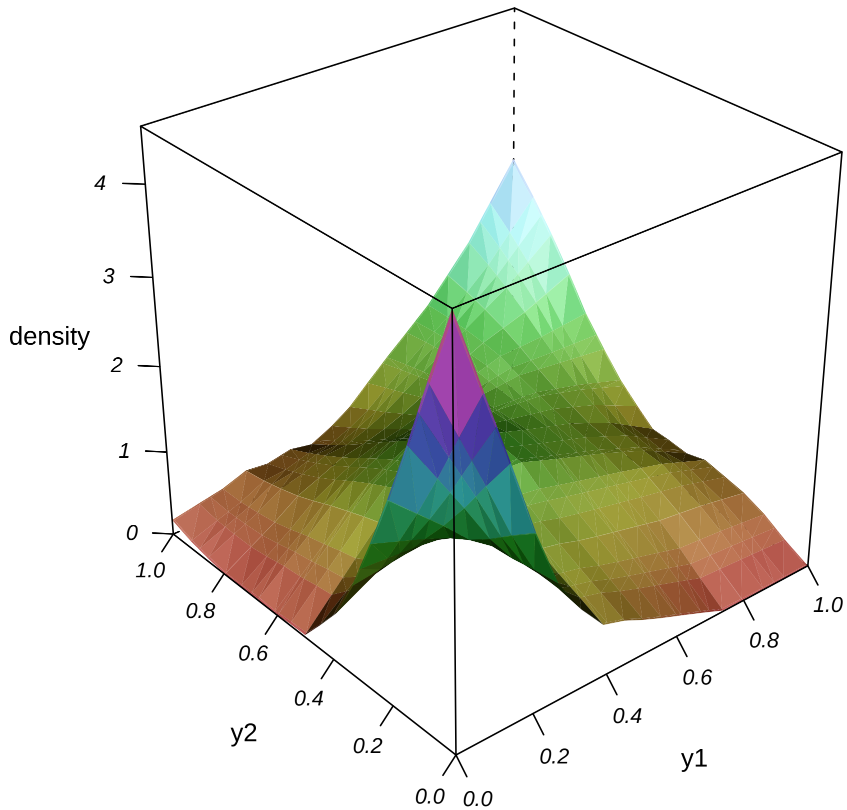
Example

Daily returns in 2006/2007 from Lufthansa and Deutsche Bank





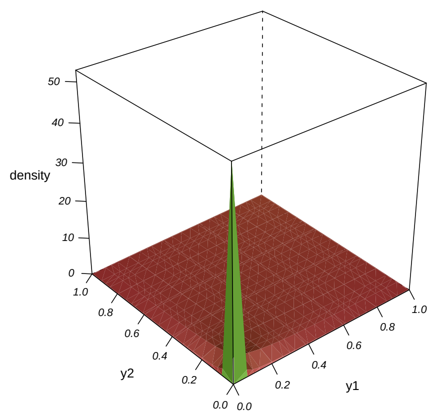
Example



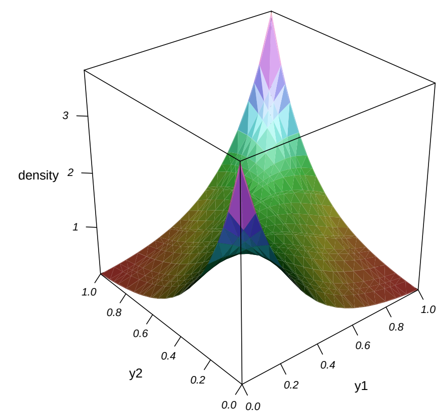
Lufthansa / Deutsche Bank return 2006/07

Example

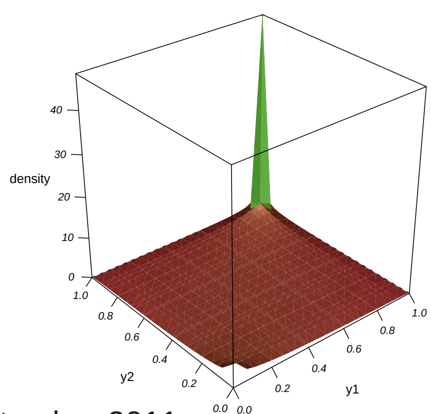
Clayton: loglik = 37.46



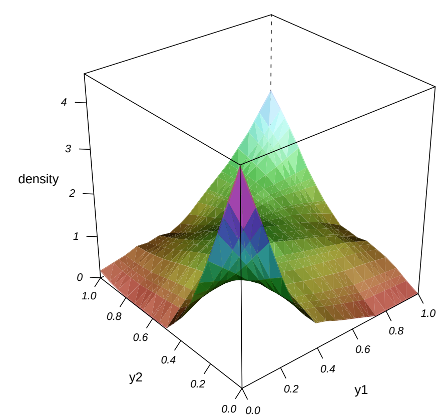
Frank: loglik = 38.80



Gumbel: loglik = 31.63



B-spline: loglik = 49.45



Tackling the Curse of Dimensionality

with Pair-Copulas

Copulas and Conditional Distributions

With Sklar's theorem we get

$$f(x_2|x_1) = c\Big(F_1(x_1), F_2(x_2)\Big) \times f_2(x_2)$$

$\Rightarrow$ We can express conditional densities with copulas.

Factorization allows to write

$$\begin{aligned} f(x_1, x_2, \ldots, x_p) \quad &= \quad f_1(x_1) \times f(x_2|x_1) \\ &\quad\quad \times f(x_3|x_1, x_2) \times \ldots \\ &\quad\quad \times f(x_p|x_1, \ldots x_{p-1}) \end{aligned}$$

The idea of Pair-Copulas

With $A = \{x_1, \ldots x_{p-2}\}$ each conditional distribution is now written as

$$f(x_p|x_{p-1}, A) = c\Big(F_p(x_p|A), F_{p-1}(x_{p-1}|A) \mid A\Big) \times f_p(x_p|A)$$

With Pair-Copulas we assume

$$c\big(F_p(x_p|A), F_{p-1}(x_{p-1}|A) \mid \underset{\substack{\uparrow \\ \text{Conditioning set}}}{A}\big) \quad = \quad c\big(F_p(x_p|A), F_{p-1}(x_{p-1}|A) \underset{\substack{\uparrow \\ \text{is omitted}}}{}\big)$$

Estimating Pair Copulas

In Pair-Copulas we need to estimate:

1. Bivariate copulas $c\big(u_p, u_{p-1}\big)$
2. Univariate conditional marginal distributions: $F_j(x_j|A)$, $j = 1, \ldots, p$.

$\Rightarrow$ We focus on the first point in this talk.

Nonparametric Estimation of Bivariate Copulas

As above, we use penalized estimation and assume

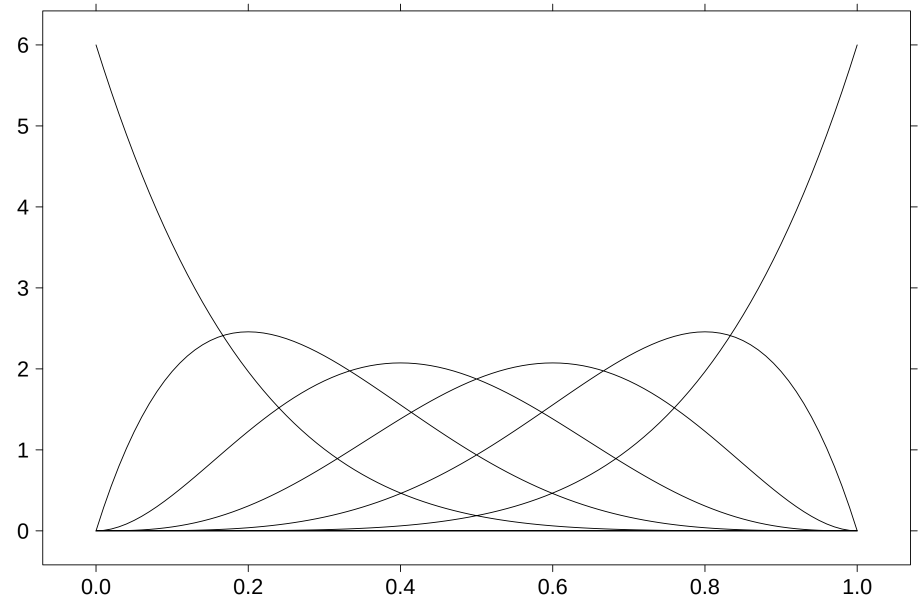
$$\begin{aligned} c(u_p, u_{p-1}) &= \sum_{k_1=0}^{K} \sum_{k_2=0}^{K} \phi_{k_1}(u_p)\phi_{k_2}(u_{p-1}) b_{k_1 k_2} \\ &= \{\phi_K(u_p) \otimes \phi_K(u_{p-1})\}\mathbf{b} \end{aligned}$$

with $\phi()$ as basis of densities on $[0, 1]$.

Bernstein Polynomials / Beta Distribution

We use Bernstein polynomials (which are Beta densities) as bases, i.e.

$$\phi_k(u) = (K+1)\binom{K}{k} u^k (1-u)^{K-k}$$



Penalization

We penalize the squared, second order derivative:

$$\int \left(\frac{\partial^2 c(u_p, u_{p-1})}{(\partial u_p)^2}\right)^2 + \left(\frac{\partial^2 c(u_p, u_{p-1})}{(\partial^2 u_{p-1})^2}\right)^2 \,\mathrm{d}u_p \,\mathrm{d}u_{p-1} \quad = \quad \underbrace{\mathbf{b}^T\mathbf{P}\mathbf{b}}_{\text{quadratic form}}$$

Assuming for the quadratic penalty

$$\mathbf{b} \sim N(0, \lambda^{-1}\mathbf{P}^{-})$$

we can again estimate λ as parameter.

Linear Constraints

- $\sum_{k_1} \phi_{k_1}(u_1) b_{k_1 \bullet} = 1$

  $\Rightarrow$ Margins are uniform.

- $\sum_{k_1, k_2} b_{k_1 k_2} = 1$

  $\Rightarrow$ The density $c()$ integrates to 1.

- $c\big(u_p, u_{p-1}\big) \geq 0$

  $\Rightarrow$ The resulting fit is a density

This is again accommodated using quadratic programming.

Estimating Univariate Conditional Margins

For Copulas one can show that:

$$
\begin{aligned}
F(x_1|x_2) &= \frac{\partial C\Big(F_1(x_1), F_2(x_2)\Big)}{\partial F_2(x_2)} \\
&= \sum_{k_1=0}^{K} \sum_{k_2=0}^{K} \Phi_{k_1}(u_1)\phi_{k_2}(u_2)b_{k_1 k_2}
\end{aligned}
$$

with $\Phi()$ as beta distribution function and $\phi()$ as beta density.

$\Rightarrow$ The Bernstein approach easily allows to calculate univariate conditional distributions.

Example

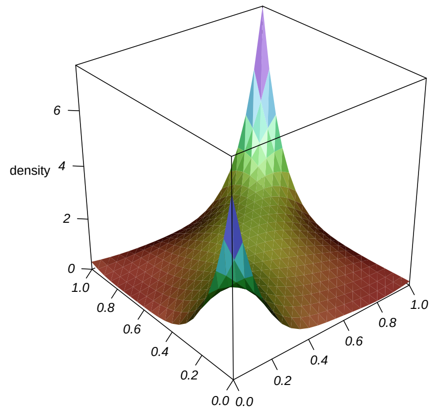
We look at the exchange rate of USD to Euro ($\mathtt{EUR}$), British Pound ($\mathtt{GBP}$) and Singapore Dollar ($\mathtt{SIN}$).

We model

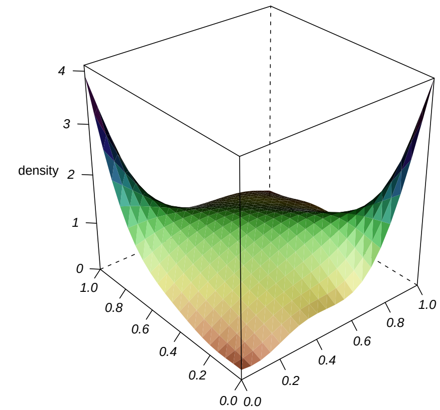
$$f(\mathtt{EUR}, \mathtt{GBP}, \mathtt{SIN}) \quad = \quad f(\mathtt{EUR}) \times f(\mathtt{GBP}|\mathtt{EUR}) \times f(\mathtt{SIN}|\mathtt{GBP}, \mathtt{EUR})$$

Example

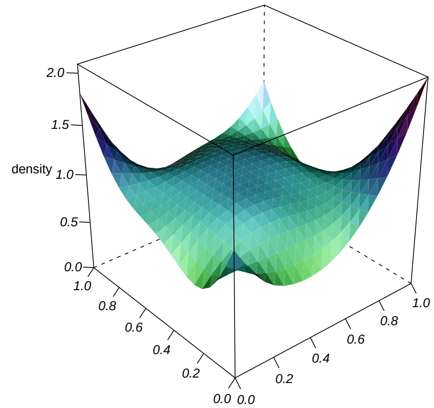
Euro against GBP



GBP against SIN



Euro against SIN given GPB



Discussion

- Penalized Spline are a flexible modelling tool.
- The link to Mixed Models allows for new, innovative statistical modelling.
- Penalized Estimation easily extends to Copula estimation.
- Quadratic programming is a useful alternative to classical Newton Raphson.
- Penalization guarantees smoothnes.
- ...

Thank you for your attention