Estimation of Copula Models with Discrete Margins *(via Bayesian Data Augmentation)*

Michael S. Smith

Melbourne Business School,

University of Melbourne

(Joint with Mohamad Khaled,

University of Queensland)

Introduction

- Copula models with discrete margins
- Distribution augmented with latent variables
- Augmented likelihood & some conditional posteriors
- Two MCMC sampling schemes for estimation; outline just one.
- Application to small online retail example
- Application to D-vine; illustration with longitudinal count data

Discrete-Margined Copula Models

- Let *X* be a vector of *m* discrete-valued random variables
- Many existing multivariate models for discrete data can be written in copula form with distribution function:

$$F(x)=C(F_1(x_1),\ldots,F_m(x_m))$$

Discrete-Margined Copula Models

- Let X be a vector of m discrete-valued random variables
- Many existing multivariate models for discrete data can be written in copula form with distribution function:

$$F(x)=C(F_1(x_1),\ldots,F_m(x_m))$$

Joint CDF of $X=(X_1,\ldots,X_m)$

Discrete-Margined Copula Models

- Let X be a vector of m discrete-valued random variables
- Many existing multivariate models for discrete data can be written in copula form with distribution function:

$$F(x)=C(F_1(x_1),\ldots,F_m(x_m))$$

Univariate CDFs of $X_1,\ldots,X_m$

Discrete-Margined Copula Models

- Let *X* be a vector of *m* discrete-valued random variables
- Many existing multivariate models for discrete data can be written in copula form with distribution function:

$$F(x)=C(F_1(x_1),\ldots,F_m(x_m))$$

Copula Function on $[0,1]^m$

Discrete-Margined Copula Models

- Let X be a vector of m discrete-valued random variables
- Many existing multivariate models for discrete data can be written in copula form with distribution function:

$$F(x)=C(F_1(x_1),\ldots,F_m(x_m))$$

- For arbitrary F, the copula function C is not unique
- Nevertheless, F is a well-defined distribution function when C is a parametric copula function

Discrete-Margined Copula Models

- We use the differencing notation:

$$\Delta_{a_k}^{b_k} C(u_1,\ldots,u_{k-1},v_k,u_{k+1},\ldots,u_m) =$$

$$C(u_1,\ldots,u_{k-1},b_k,u_{k+1},\ldots,u_m) - C(u_1,\ldots,u_{k-1},a_k,u_{k+1},\ldots,u_m)$$

Discrete-Margined Copula Models

- We use the differencing notation:

$$\Delta_{a_k}^{b_k} C(u_1, \ldots, u_{k-1}, v_k, u_{k+1}, \ldots, u_m) =$$

$$C(u_1, \ldots, u_{k-1}, b_k, u_{k+1}, \ldots, u_m) - C(u_1, \ldots, u_{k-1}, a_k, u_{k+1}, \ldots, u_m)$$

- The v_k is simply an “index of differencing”

Discrete-Margined Copula Models

- We use the differencing notation:

$$\Delta_{a_k}^{b_k} C(u_1,\ldots,u_{k-1},v_k,u_{k+1},\ldots,u_m) =$$
$$C(u_1,\ldots,u_{k-1},b_k,u_{k+1},\ldots,u_m) - C(u_1,\ldots,u_{k-1},a_k,u_{k+1},\ldots,u_m)$$

- In that case the PMF is given by

$$f(x) = \Delta_{a_1}^{b_1} \Delta_{a_2}^{b_2} \ldots \Delta_{a_m}^{b_m} C(v_1,v_2,\ldots,v_m)$$

- where

$$b_j = F_j(x_j) \qquad a_j = F_j(x_j^-)$$

Discrete-Margined Copula Models

- We use the differencing notation:

$$\Delta_{a_k}^{b_k} C(u_1,\ldots,u_{k-1},v_k,u_{k+1},\ldots,u_m) =$$
$$C(u_1,\ldots,u_{k-1},b_k,u_{k+1},\ldots,u_m) - C(u_1,\ldots,u_{k-1},a_k,u_{k+1},\ldots,u_m)$$

- In that case the PMF is given by

$$f(x) = \Delta_{a_1}^{b_1} \Delta_{a_2}^{b_2} \ldots \Delta_{a_m}^{b_m} C(v_1, v_2, \ldots, v_m)$$

- where

Left-hand Limit at x_j

$$b_j = F_j(x_j) \qquad a_j = F_j(x_j^-)$$

Discrete-Margined Copula Models

- We use the differencing notation:

$$\Delta_{a_k}^{b_k} C(u_1,\ldots,u_{k-1},v_k,u_{k+1},\ldots,u_m) =$$

$$C(u_1,\ldots,u_{k-1},b_k,u_{k+1},\ldots,u_m) - C(u_1,\ldots,u_{k-1},a_k,u_{k+1},\ldots,u_m)$$

- In that case the PMF is given by

$$f(x) = \Delta_{a_1}^{b_1} \Delta_{a_2}^{b_2} \ldots \Delta_{a_m}^{b_m} C(v_1,v_2,\ldots,v_m)$$

- where

For ordinal data

$$b_j = F_j(x_j) \qquad a_j = F_j(x_j^-) = F_j(x_j - 1)$$

Difficulties with Estimation

- Genest & Nešlehová (07) highlight the problems of using rank-based estimators
- However, in general, it is difficult to compute MLE of the copula parameters because:
  - evaluation of the PMF (and hence MLE) involves $O(2^m)$ computations
  - Direct maximization of the likelihood can be difficult

Augmented Distribution

- To circumnavigate both problems, we consider augmenting the distribution of *X* with $U=(U_1,\ldots,U_m)$ so that

$$f(x_j|u_j) = I\left(F_j(x_j) \leq u_j < F_j(x_j^-)\right)$$

- where:
  - $\mathcal{I}(A)=1$ if *A* is true, and $\mathcal{I}(A)=0$ if *A* is false

Augmented Distribution

- To circumnavigate both problems, we consider augmenting the distribution of *X* with $U=(U_1,\ldots,U_m)$ so that

$$f(x,u) = f(x|u)f(u)$$

- where:
  - $\mathscr{I}(A)=1$ if *A* is true, and $\mathscr{I}(A)=0$ if *A* is false

Augmented Distribution

- To circumnavigate both problems, we consider augmenting the distribution of X with $U=(U_1,\ldots,U_m)$ so that

$$f(x,u) = f(x\,|\,u)c(u) = \prod_{j=1}^{m} \mathcal{I}(F_j(x_j^-) \leq u_j < F_j(x_j))c(u)$$

- where:
  - $\mathcal{I}(A)=1$ if A is true, and $\mathcal{I}(A)=0$ if A is false
  - $c(u)=\partial C(u)/\partial u$ is the copula density for C
- This is a "mixed augmented density"

Augmented Distribution

- To circumnavigate both problems, we consider augmenting the distribution of X with $U=(U_1,\ldots,U_m)$ so that

$$f(x,u) = f(x\,|\,u)c(u) = \prod_{j=1}^{m} \mathcal{I}(F_j(x_j^-) \leq u_j < F_j(x_j))c(u)$$

- It can be shown that the <u>*marginal PMF of X*</u> is that of the copula model
- The aim is to construct <u>*likelihood-based*</u> inference using the <u>*augmented posterior*</u> constructed using $f(x,u)$

Latent Variable Distributions

- In our DA approach we sample the *U*'s explicitly
- The latent variable *U* (conditional on *X*) follows a multivariate constrained distribution

$$f(u \mid x) = \frac{c(u)}{f(x)} \prod_{j=1}^{m} \mathcal{I}(a_j \leq u_j < b_j)$$

Two MCMC DA Schemes

- *Scheme 1:*
  - Generates *u* as a block using MH with an approximation *q(u)* which is “close to” *f(u|x)*
  - Need to compute the conditional copula CDFs $C_{j|1,\ldots,j-1}$ a total of *5(m-1)* times
- *Scheme 2:*
  - Generates u_j one-at-a-time
  - Need to compute the conditional copula CDFs $C_{j|k\neq j}$ a total of *m* times
- Can use at least one scheme for all copula models currently being employed

Latent Variable Distributions

•The development of Scheme 1 relies on the derivation of the following *conditional distribution*

$$f(u_j \mid u_1,\ldots,u_{j-1},x) =$$

$$c_{j|1,\ldots,j-1}(u_j \mid u_1,\ldots,u_{j-1})\,\mathcal{I}(a_j \le u_j < b_j)\,\mathcal{K}_j(u_1,\ldots,u_j)$$

Conditional copula density

Latent Variable Distributions

- The development of Scheme 1 relies on the derivation of the following *<u>conditional distribution</u>*

$$f(u_j \mid u_1,\ldots,u_{j-1},x) =$$

$$c_{j|1,\ldots,j-1}(u_j \mid u_1,\ldots,u_{j-1})\,\mathcal{I}(a_j \le u_j < b_j)\,\mathcal{K}_j(u_1,\ldots,u_j)$$

Constrained to $[a_j,b_j)$

Latent Variable Distributions

- The development of Scheme 1 relies on the derivation of the following <u>*conditional distribution*</u>

$$f(u_j \mid u_1,\ldots,u_{j-1},x) =$$

$$c_{j|1,\ldots,j-1}(u_j \mid u_1,\ldots,u_{j-1})\,\mathcal{I}(a_j \le u_j < b_j)\,\mathcal{K}_j(u_1,\ldots,u_j)$$

With a $O(2^{m-j})$ term that is a function of $u_1,\ldots,u_j$

Generating *u*: the MH Proposal

- The proposal density for *u* is:

$$g_j(u) = \prod_{j=2}^{m} g_j(u_j \mid u_1, \ldots, u_{j-1}) g_1(u_1)$$

- Generate sequentially from each g_j $(j=1,\ldots,m)$

Generating u: the MH Proposal

- The proposal density for u is:

$$g_j(u) = \prod_{j=2}^{m} g_j(u_j \mid u_1, \ldots, u_{j-1}) g_1(u_1)$$

- where:

$$g_j(u_j \mid \text{---}) = \frac{c_{j|1,\ldots,j-1}(u_j \mid \text{---}; \varphi)\mathscr{I}(a_j \le u_j < b_j)}{C_{j|1,\ldots,j-1}(b_j \mid \text{---}; \varphi) - C_{j|1,\ldots,j-1}(a_j \mid \text{---}; \varphi)}$$

- and

$$g_1(u_{i1}) = \mathscr{I}(a_{i1} \le u_{i1} < b_{i1}) / (b_{1j} - a_{1j})$$

Generating *u*: the MH Proposal

- The proposal density for *u* is:

$$g_j(u) = \prod_{j=2}^{m} g_j(u_j \mid u_1, \ldots, u_{j-1})\, g_1(u_1)$$

- where:

$$g_j(u_j \mid \text{--}) = \frac{c_{j|1,\ldots,j-1}(u_j \mid \text{--}; \varphi)\,\mathscr{I}(a_j \le u_j < b_j)}{C_{j|1,\ldots,j-1}(b_j \mid \text{--}; \varphi) - C_{j|1,\ldots,j-1}(a_j \mid \text{--}; \varphi)}$$

- and

$$g_1(u_{i1}) = \mathscr{I}(a_{i1} \le u_{i1} < b_{i1}) \,/\, (b_{1j} - a_{1j})$$

Just saving space with this notation!

Generating *u*: the MH Proposal

- The proposal density for *u* is:

$$g_j(u) = \prod_{j=2}^{m} g_j(u_j \mid u_1, \ldots, u_{j-1}) g_1(u_1)$$

- where:

$$g_j(u_j \mid --) = \frac{c_{j|1,\ldots,j-1}(u_j \mid --; \varphi)\mathcal{I}(a_j \le u_j < b_j)}{C_{j|1,\ldots,j-1}(b_j \mid --; \varphi) - C_{j|1,\ldots,j-1}(a_j \mid --; \varphi)}$$

- and

$$g_1(u_{i1}) = \mathcal{I}(a_{i1} \le u_{i1} < b_{i1}) \,/\, (b_{1j} - a_{1j})$$

Constrained conditional copula distribution

Generating *u*: the MH Proposal

- The proposal density for *u* is:

$$g_j(u) = \prod_{j=2}^{m} g_j(u_j \mid u_1,\ldots,u_{j-1})\, g_1(u_1)$$

- where:

$$g_j(u_j \mid --) = \frac{c_{j|1,\ldots,j-1}(u_j \mid --;\varphi)\,\mathscr{I}(a_j \le u_j < b_j)}{C_{j|1,\ldots,j-1}(b_j \mid --;\varphi) - C_{j|1,\ldots,j-1}(a_j \mid --;\varphi)}$$

- and

$$g_1(u_{i1}) = \mathscr{I}(a_{i1} \le u_{i1} < b_{i1}) \,/\, (b_{1j} - a_{1j})$$

The normalising constants…

Generating *u*: the MH Proposal

- The proposal density for *u* is:

$$g_j(u) = \prod_{j=2}^{m} g_j(u_j \mid u_1, \ldots, u_{j-1}) g_1(u_1)$$

- where:

$$g_j(u_j \mid --) = \frac{c_{j|1,\ldots,j-1}(u_j \mid --; \varphi)\mathscr{I}(a_j \le u_j < b_j)}{C_{j|1,\ldots,j-1}(b_j \mid --; \varphi) - C_{j|1,\ldots j-1}(a_j \mid --; \varphi)}$$

- and

$$g_1(u_{i1}) = \mathscr{I}(a_{i1} \le u_{i1} < b_{i1}) / (b_{1j} - a_{1j})$$

To implement, just need to be able to compute $C_{j|1,\ldots j-1}$ and its inverse… *3(m-1)* times

Generating *u*: the MH Proposal

- The proposal density for *u* is:

$$g_j(u) = \prod_{j=2}^{m} g_j(u_j \mid u_1, \ldots, u_{j-1}) g_1(u_1)$$

- where:

$$g_j(u_j \mid --) = \frac{c_{j|1,\ldots,j-1}(u_j \mid --;\varphi)\mathscr{I}(a_j \le u_j < b_j)}{C_{j|1,\ldots,j-1}(b_j \mid --;\varphi) - C_{j|1,\ldots j-1}(a_j \mid --;\varphi)}$$

- and

$$g_1(u_{i1}) = \mathscr{I}(a_{i1} \le u_{i1} < b_{i1}) / (b_{1j} - a_{1j})$$

As $|F_j(x_j)\text{-}F_j(x_j^-)|\to 0$, then $g(u)\to f(u|\phi,x)$,
So that is a "close" approximation

Generating φ given u

- Conditional on *u*, it is much easier to generate any copula parameters φ
- Posterior is:

$$f(\varphi|u,\Theta,x) = f(\varphi|u)$$

Generating φ given u

- Conditional on *u*, it is much easier to generate any copula parameters φ
- Posterior is:

$$f(\varphi|u,\Theta,x) = f(\varphi|u) = \prod_{i}^{n} c(u_i|\varphi)\pi(\varphi)$$

Generating *φ* given *u*

- Conditional on *u*, it is much easier to generate any copula parameters *φ*
- Posterior is:

$$f(\varphi|u,\Theta,x) = f(\varphi|u)$$

$$= \prod_{i}^{n} c(u_i|\varphi)\pi(\varphi)$$

copula density evaluated at each vector $u_i=(u_{i1},\ldots,u_{im})'$

Generating φ given u

- Conditional on u, it is much easier to generate any copula parameters φ
- Posterior is:

$$f(\varphi|u,\Theta,x) = f(\varphi|u)$$
$$= \prod_{i}^{n} c(u_i|\varphi)\pi(\varphi)$$

prior structure

Bayesian Estimation: Advantages

- Provides likelihood-based inference (particularly important for this model)
- Can compute dependence structure of *U*, and of *X*, from fitted copula model
- Allows for shrinkage priors, such as:
  - for correlation matrix (eg Pitt et al. 06; Daniels & Pourahmadi 09)
  - model averaging (Smith et al. 10/Czado & Min'11)
  - hierarchical models (eg. Almeida & Czado '10)
- Numerically robust

Illustration: Online Retail

- $n=10{,}000$ randomly selected visits to amazon.com collected by ComScore
- Bivariate example with:
  - $X_1 \in \{1,2,3,\ldots\}$ = # of unique page views
  - $X_2 \in \{0,1\}$ = sales incidence
- 92% of observations are non-zeros
- Positive dependence between X_1 and X_2
- Three different bivariate copulas with positive dependence:
  - Clayton, BB7, Gaussian

Illustration: Online Retail

	Bayes	MLE	PMLE
Clayton Copula			
$\hat{\phi}$	4.960	5.099	0.838
	(4.616, 5.309)	(0.182)	(0.020)
$\hat{\tau}$	0.713	0.718	0.293
	(0.698, 0.726)	(0.007)	(0.005)
$\hat{\lambda}^L$	0.869	0.873	0.437
	(0.861, 0.878)	(0.004)	(0.009)
$\hat{\tau}^F$	0.1056	0.1055	–
	(0.1037, 0.1072)	(0.0010)	–
BB7 Copula			
$\hat{\phi}_1$	1.008	1.000	1.000
	(1.000, 1.026)	(0.030)	(0.001)
$\hat{\phi}_2$	4.972	5.095	0.837
	(4.589, 5.308)	(0.183)	(0.020)
$\hat{\tau}$	0.713	0.718	0.295
	(0.696, 0.726)	(0.007)	(0.005)
$\hat{\lambda}^L$	0.870	0.873	0.440
	(0.860, 0.878)	(0.004)	(0.009)
$\hat{\lambda}^U$	0.011	0.000	0.000
	(0.000, 0.034)	(0.041)	(0.001)
$\hat{\tau}^F$	0.1048	0.1055	–
	(0.1042, 0.1055)	(0.0013)	–
Gaussian Copula			
$\hat{\phi}$	0.635	0.637	0.128
	(0.506, 0.738)	(0.068)	(0.027)
$\hat{\tau}$	0.440	0.440	0.081
	(0.337, 0.528)	(0.056)	(0.017)
$\hat{\tau}^F$	0.0983	0.0990	–
	(0.0806, 0.1128)	(0.0096)	–

Illustration: Online Retail

	Bayes	MLE	PMLE
		Clayton Copula	
$\hat{\phi}$	4.960	5.099	0.838
	(4.616, 5.309)	(0.182)	(0.020)
$\hat{\tau}$	0.713	0.718	0.293
	(0.698, 0.726)	(0.007)	(0.005)
$\hat{\lambda}^L$	0.869	0.873	0.437
	(0.861, 0.878)	(0.004)	(0.009)
$\hat{\tau}^F$	0.1056	0.1055	–
	(0.1037, 0.1072)	(0.0010)	–
		BB7 Copula	
$\hat{\phi}_1$	1.008	1.000	1.000
	(1.000, 1.026)	(0.030)	(0.001)
$\hat{\phi}_2$	4.972	5.095	0.837
	(4.589, 5.308)	(0.183)	(0.020)
$\hat{\tau}$	0.713	0.718	0.295
	(0.696, 0.726)	(0.007)	(0.005)
$\hat{\lambda}^L$	0.870	0.873	0.440
	(0.860, 0.878)	(0.004)	(0.009)
$\hat{\lambda}^U$	0.011	0.000	0.000
	(0.000, 0.034)	(0.041)	(0.001)
$\hat{\tau}^F$	0.1048	0.1055	–
	(0.1042, 0.1055)	(0.0013)	–
		Gaussian Copula	
$\hat{\phi}$	0.635	0.637	0.128
	(0.506, 0.738)	(0.068)	(0.027)
$\hat{\tau}$	0.440	0.440	0.081
	(0.337, 0.528)	(0.056)	(0.017)
$\hat{\tau}^F$	0.0983	0.0990	–
	(0.0806, 0.1128)	(0.0096)	–

Bayes same as MLE: reassuring

Illustration: Online Retail

	Bayes	MLE	PMLE
	Clayton Copula		
$\hat{\phi}$	4.960	5.099	0.838
	(4.616, 5.309)	(0.182)	(0.020)
$\hat{\tau}$	0.713	0.718	0.293
	(0.698, 0.726)	(0.007)	(0.005)
$\hat{\lambda}^L$	0.869	0.873	0.437
	(0.861, 0.878)	(0.004)	(0.009)
$\hat{\tau}^F$	0.1056	0.1055	–
	(0.1037, 0.1072)	(0.0010)	–
	BB7 Copula		
$\hat{\phi}_1$	1.008	1.000	1.000
	(1.000, 1.026)	(0.030)	(0.001)
$\hat{\phi}_2$	4.972	5.095	0.837
	(4.589, 5.308)	(0.183)	(0.020)
$\hat{\tau}$	0.713	0.718	0.295
	(0.696, 0.726)	(0.007)	(0.005)
$\hat{\lambda}^L$	0.870	0.873	0.440
	(0.860, 0.878)	(0.004)	(0.009)
$\hat{\lambda}^U$	0.011	0.000	0.000
	(0.000, 0.034)	(0.041)	(0.001)
$\hat{\tau}^F$	0.1048	0.1055	–
	(0.1042, 0.1055)	(0.0013)	–
	Gaussian Copula		
$\hat{\phi}$	0.635	0.637	0.128
	(0.506, 0.738)	(0.068)	(0.027)
$\hat{\tau}$	0.440	0.440	0.081
	(0.337, 0.528)	(0.056)	(0.017)
$\hat{\tau}^F$	0.0983	0.0990	–
	(0.0806, 0.1128)	(0.0096)	–

Psuedo MLE is total junk

Illustration: Online Retail

	Bayes	MLE	PMLE
	Clayton Copula		
$\hat{\phi}$	4.960	5.099	0.838
	(4.616, 5.309)	(0.182)	(0.020)
$\hat{\tau}$	0.713	0.718	0.293
	(0.698, 0.726)	(0.007)	(0.005)
$\hat{\lambda}^L$	0.869	0.873	0.437
	(0.861, 0.878)	(0.004)	(0.009)
$\hat{\tau}^F$	0.1056	0.1055	–
	(0.1037, 0.1072)	(0.0010)	–
	BB7 Copula		
$\hat{\phi}_1$	1.008	1.000	1.000
	(1.000, 1.026)	(0.030)	(0.001)
$\hat{\phi}_2$	4.972	5.095	0.837
	(4.589, 5.308)	(0.183)	(0.020)
$\hat{\tau}$	0.713	0.718	0.295
	(0.696, 0.726)	(0.007)	(0.005)
$\hat{\lambda}^L$	0.870	0.873	0.440
	(0.860, 0.878)	(0.004)	(0.009)
$\hat{\lambda}^U$	0.011	0.000	0.000
	(0.000, 0.034)	(0.041)	(0.001)
$\hat{\tau}^F$	0.1048	0.1055	–
	(0.1042, 0.1055)	(0.0013)	–
	Gaussian Copula		
$\hat{\phi}$	0.635	0.637	0.128
	(0.506, 0.738)	(0.068)	(0.027)
$\hat{\tau}$	0.440	0.440	0.081
	(0.337, 0.528)	(0.056)	(0.017)
$\hat{\tau}^F$	0.0983	0.0990	–
	(0.0806, 0.1128)	(0.0096)	–

Kendall's tau for $U \in [0,1]^m$ differs from Kendall's tau for X

Illustration: Online Retail

	Bayes	MLE	PMLE
Clayton Copula			
$\hat{\phi}$	4.960	5.099	0.838
	(4.616, 5.309)	(0.182)	(0.020)
$\hat{\tau}$	0.713	0.718	0.293
	(0.698, 0.726)	(0.007)	(0.005)
$\hat{\lambda}^L$	0.869	0.873	0.437
	(0.861, 0.878)	(0.004)	(0.009)
$\hat{\tau}^F$	0.1056	0.1055	–
	(0.1037, 0.1072)	(0.0010)	–
BB7 Copula			
$\hat{\phi}_1$	1.008	1.000	1.000
	(1.000, 1.026)	(0.030)	(0.001)
$\hat{\phi}_2$	4.972	5.095	0.837
	(4.589, 5.308)	(0.183)	(0.020)
$\hat{\tau}$	0.713	0.718	0.295
	(0.696, 0.726)	(0.007)	(0.005)
$\hat{\lambda}^L$	0.870	0.873	0.440
	(0.860, 0.878)	(0.004)	(0.009)
$\hat{\lambda}^U$	0.011	0.000	0.000
	(0.000, 0.034)	(0.041)	(0.001)
$\hat{\tau}^F$	0.1048	0.1055	–
	(0.1042, 0.1055)	(0.0013)	–
Gaussian Copula			
$\hat{\phi}$	0.635	0.637	0.128
	(0.506, 0.738)	(0.068)	(0.027)
$\hat{\tau}$	0.440	0.440	0.081
	(0.337, 0.528)	(0.056)	(0.017)
$\hat{\tau}^F$	0.0983	0.0990	–
	(0.0806, 0.1128)	(0.0096)	–

Clayton and BB7 copulas identify strong lower tail dependence in the u-space…..

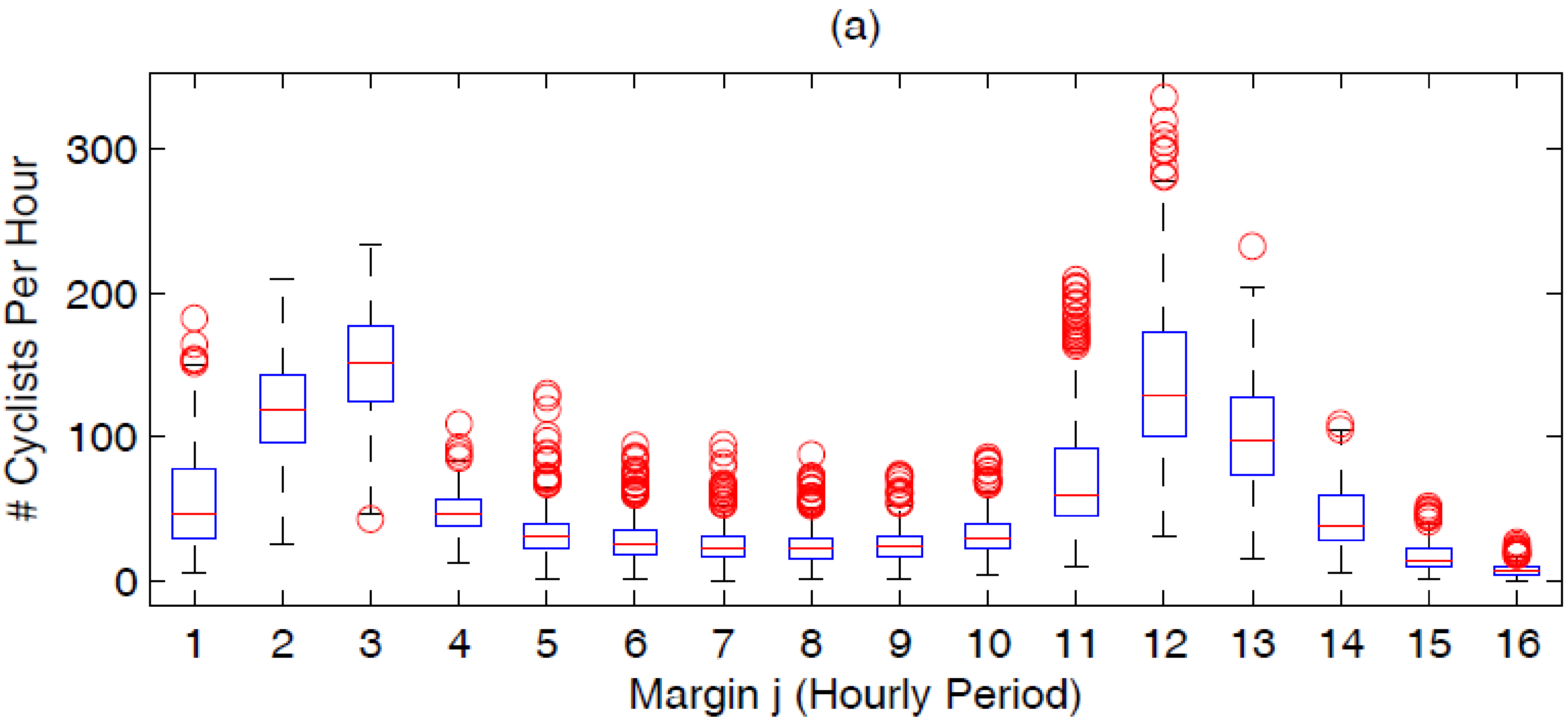
Illustration: (Parsimonious) D-vine for Bicycle Counts

- *Longitudinal* count data where:

  X_{ij} = # of bicycles on working day i during hour j

- Collected on an off-road bike path in Melbourne used for commuting
- Counts highly variable due to high variance in weather conditions
- $m=16$, $n=565$
- Use EDFs for the margins, and D-vine for C (with selection of independence pair-cops.)

Counts



(a)
Cyclists Per Hour
300
200
100
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Margin j (Hourly Period)

D-vine

- The vector $X=(X_1,\ldots,X_{16})$ is longitudinal
- A D-vine is a particularly good choice for the dependence structure when the process is likely to exhibit *Markov structure*
- Note that from Smith et al. (10) in a D-vine:

$$C_{j|1,\ldots,j-1}(u_j|u_1,\ldots,u_{j-1}) = h_{j,1} \circ h_{j,2} \circ \ldots \circ h_{j,j-1}(u_j)$$

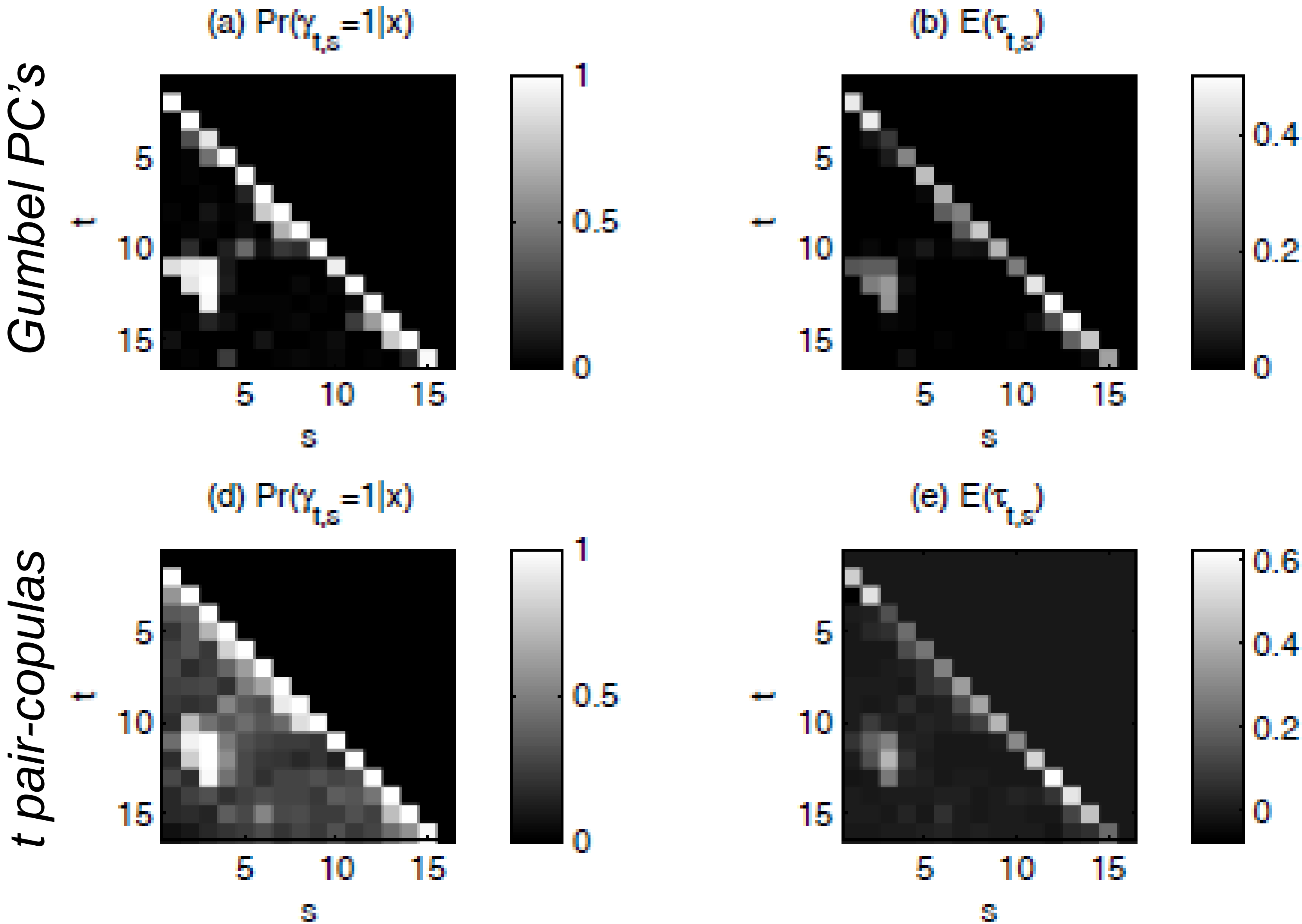
$$C_{j|1,\ldots,j-1}^{-1}(z_j|u_1,\ldots,u_{j-1}) = h_{j,j-1}^{-1} \circ h_{j,j-2}^{-1} \circ \ldots \circ h_{j,1}^{-1}(z_j)$$

- The $h_{j,t}$ functions are the conditional CDFs of the pair-copulas (see Joe 96; Aas et al. 09 and others)

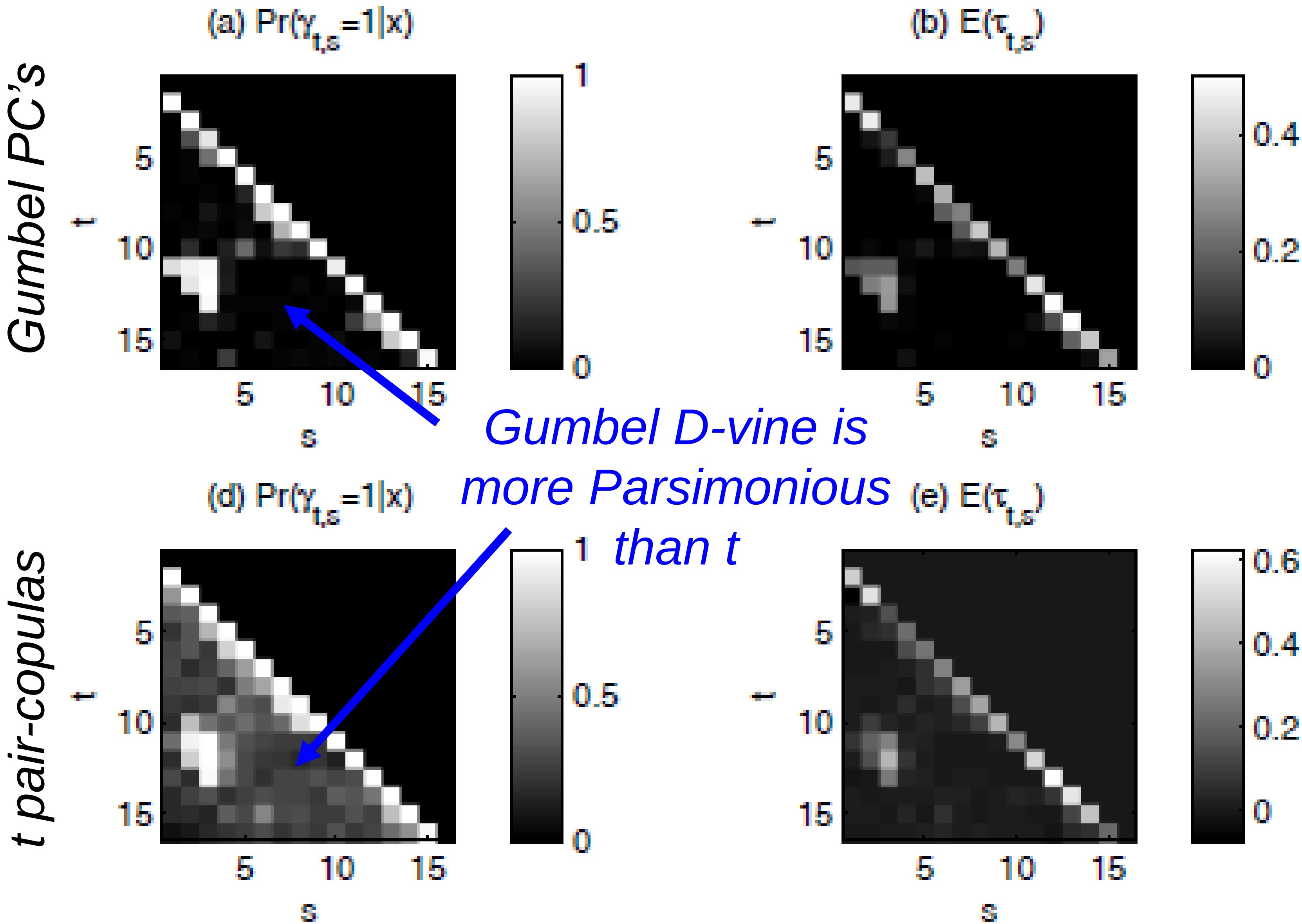
D-vine: Models

- We use three D-vines with “pair-copula selection” and:
  - *Gumbel pair-copulas*
  - *Clayton pair-copulas*
  - *t pair-copulas (two parameter copula)*
- Some objectives are to see:
  - Whether there is parsimony in the D-vines?
  - Whether choice of pair-copula type makes a difference?
  - Can you predict the evening peak *($j=12$)* given the morning peak *($j=3$)?*

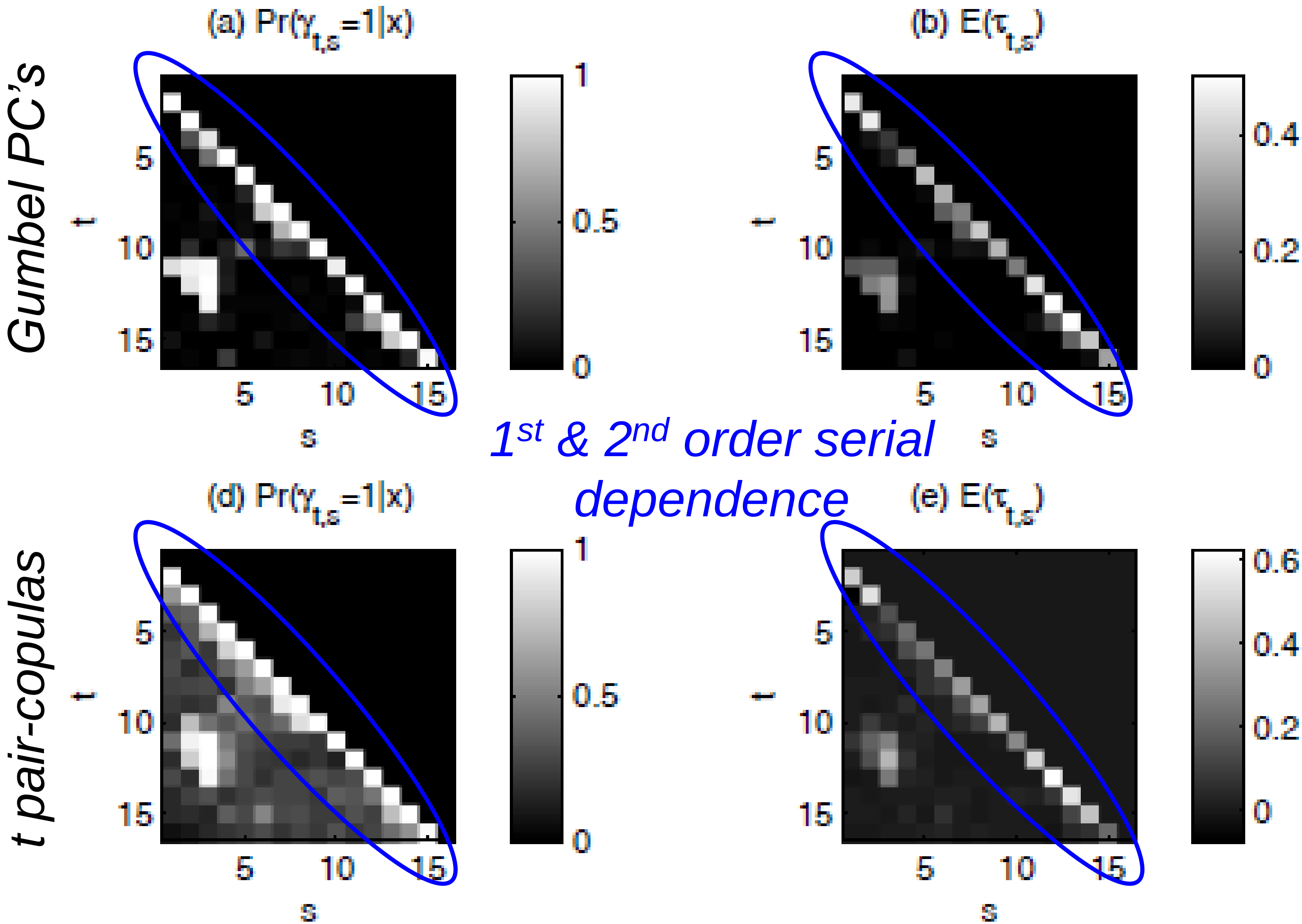
(m(m-1)/2) Pair-Copula Estimates



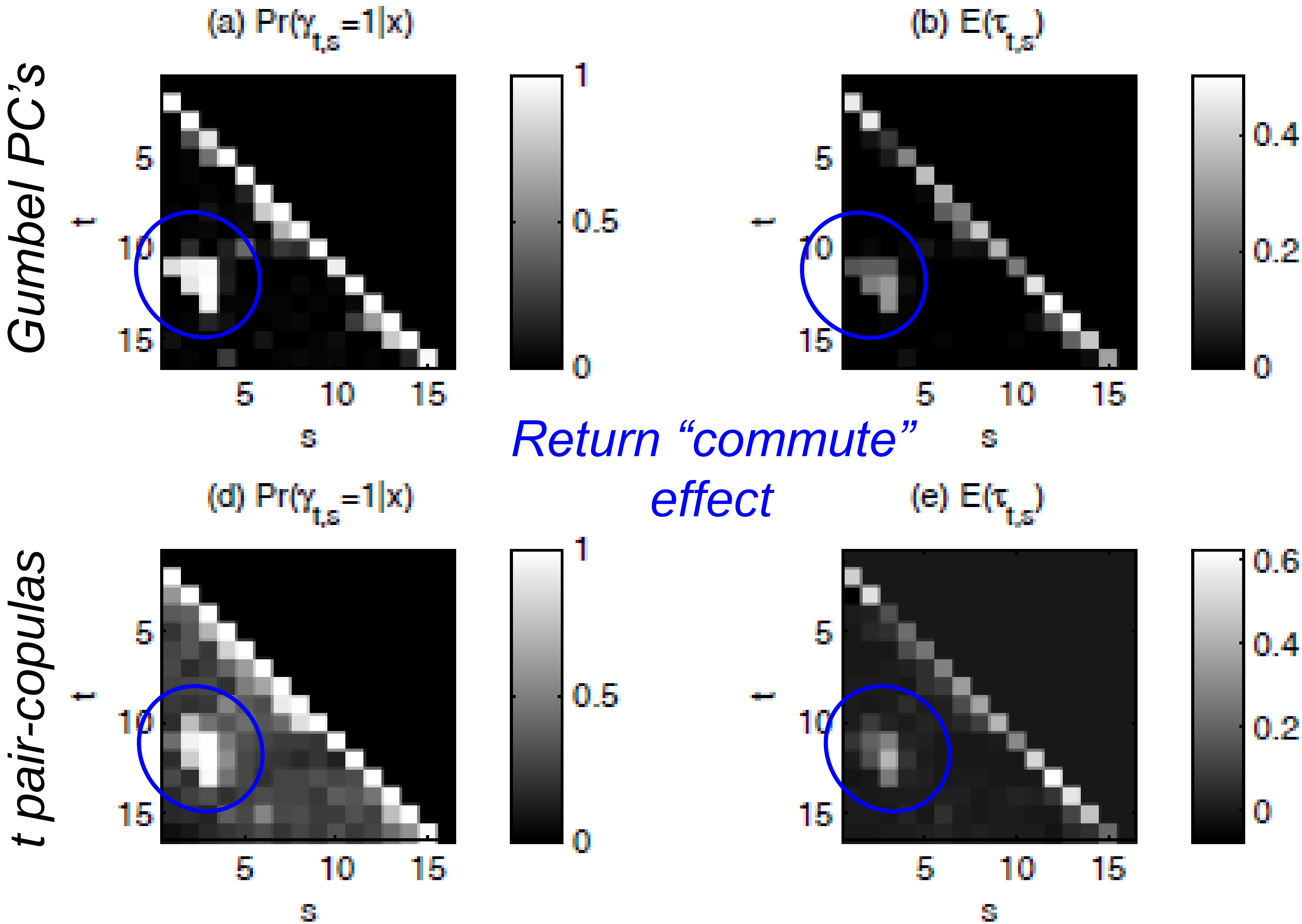
(m(m-1)/2) Pair-Copula Estimates



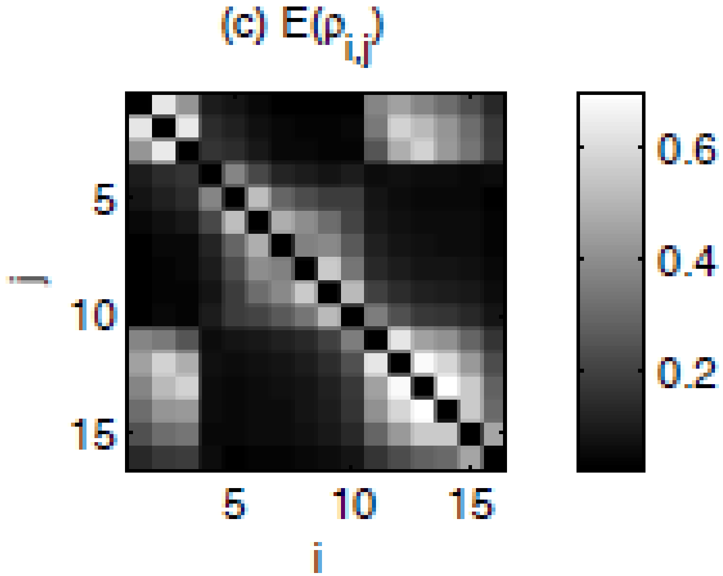
(m(m-1)/2) Pair-Copula Estimates



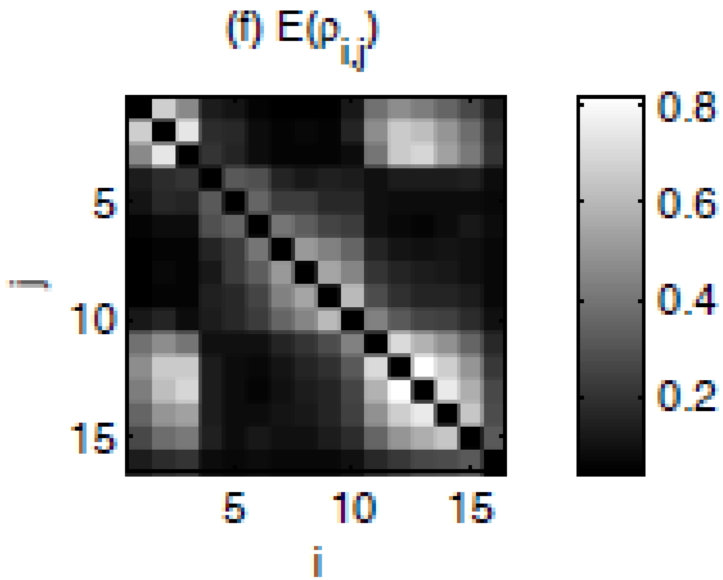
(m(m-1)/2) Pair-Copula Estimates



Spearman Pairwise Dependences



<- From the Parsimonious D-vine with ***Gumbel*** PC's



<- From the Parsimonious D-vine with ***t*** PC's

Bivariate Margins

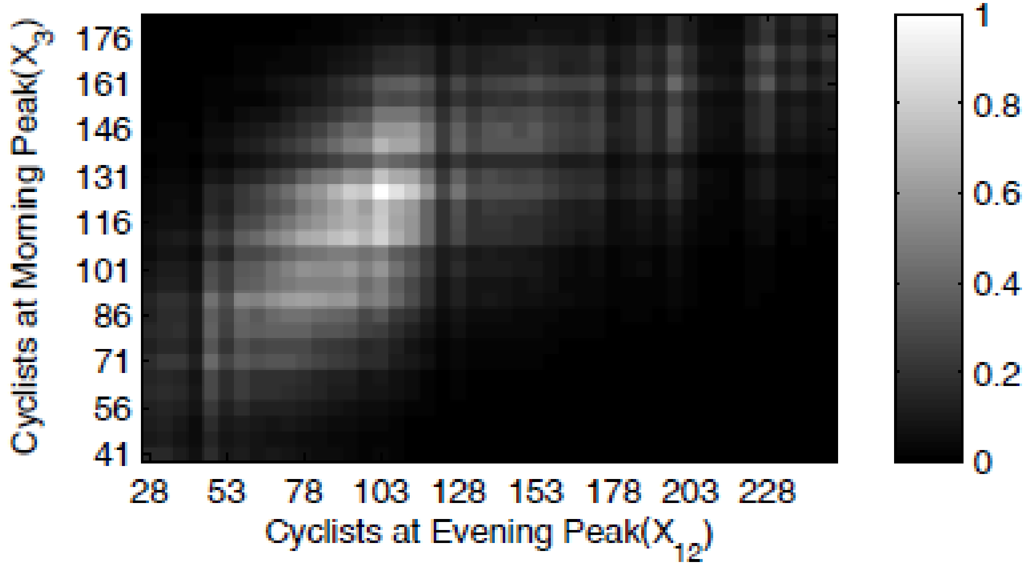
- We compute the bivariate margins in:
  - X_3: the morning peak hour on the bike path
  - X_{12}: the evening peak hour on the bike path

$$F_{3,12}(x'_3, x'_{12}) = \int C_{3,12}(F_3(x'_3), F_{12}(x'_{12}); \phi) f(\phi \mid x) d\phi$$

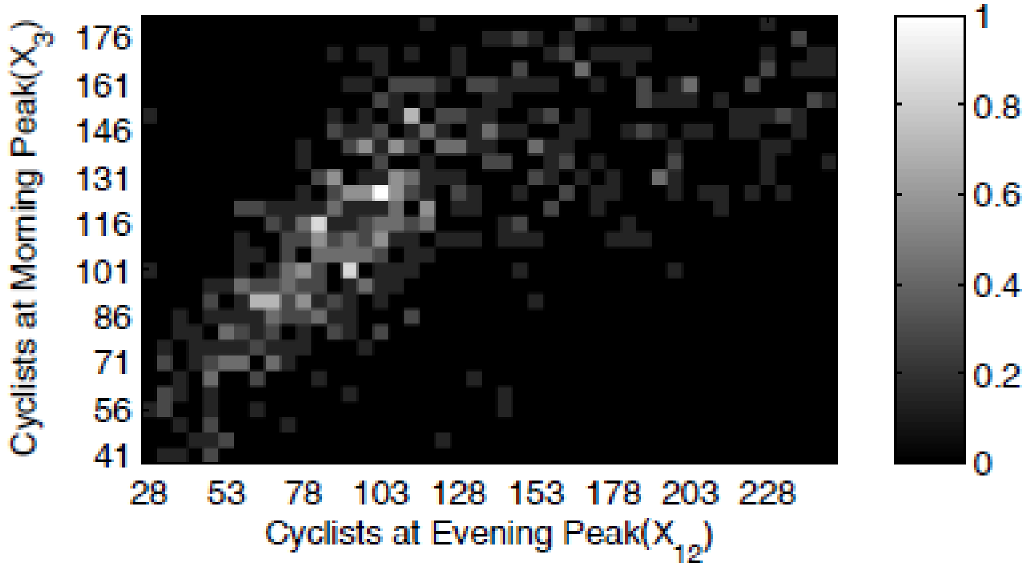
- The dependence parameter is integrated out with respect to its posterior distribution (ie "fitted" distribution)

Bivariate Margins

(b) t–copula



(c) Bivariate Data Histogram



Mixed Margins

- The approach can be extended to the case where some margins are discrete, others continuous
- Latent variables are only introduced for the discrete margins
- Extending the earlier results to this case is ***non-trivial*** (see paper)
- But once done, adjusted versions of Sampling Schemes 1 and 2 can be derived (see paper)

Some Features of Approach

- A general approach applicable to all popular parametric copula functions
- At least one of the two sampling schemes can be used for a given copula model
- Speed depends upon how fast it is to compute $C_{j|1,\ldots,j-1}$ and/or $C_{j|k\neq j}$
- It is likelihood-based; see discussion in Genest & Nešlehová (07) & Song et al. (09/10) for the importance of this

Some Features of Approach

- For copulas constructed by *inversion* of distribution *G*, probably better to augment with latents $X^* \sim G$ (cf: Pitt et al. 06; Smith, Gan & Kohn 10; Danaher & Smith 11)
- Not widely appreciated that the Gaussian copula is as restrictive for some discrete data, just as for continuous data (cf: Nikoloulopoulos & Karlis 08; 10)
- Similarly, with model averaging (eg. in a pair-copula model in Smith et al. 10)