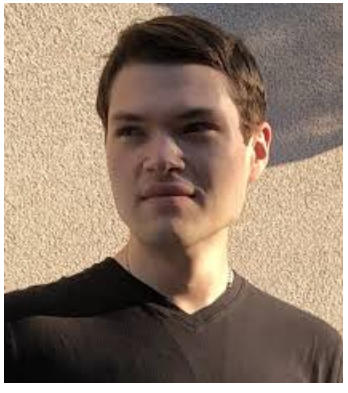
Statistical Elucidation of Latent Structures via Copulas

Radu Craiu

Department of Statistical Sciences
University of Toronto

Joint with Robert Zimmerman (Imperial College, London)



SPSR, Bucharest, May 2025

Overview of Latent Variable Models

- Latent Variable Models (LVMs) were driven by the need to model unobserved variables/putative constructs.
- Applications span psychology, sociology, economics, medicine, and machine learning.
- Key motivations include:
  - Measuring unobservable traits
  - Dealing with heterogeneity
  - Reducing dimensionality
  - Modeling dependence structure

Psychology and Educational Testing

- **Motivation**: Measure intelligence, personality, attitudes.
- **Latent Trait Models**: Spearman's g-factor, IRT models.
  - Spearman (1904): Proposed the "g-factor" (general intelligence) as a latent cause of test performance.
  - Thurstone (1935): Introduced multiple factor models for mental abilities.
  - Hotelling (1933): Principal Component Analysis (PCA) – not latent but foundational for dimensionality reduction.

Sociology and Marketing

- **Motivation**: Identify hidden subpopulations or belief systems.
- **Latent Class Analysis (LCA)**:
  - Cluster categorical responses into latent groups.
  - Anderson and Rubin (1956): Inference in factor analysis; identifiability and estimation.
  - Lazarsfeld and Henry (1968): Latent Class Analysis (LCA) for categorical data.
- **Latent Mixture Models**:
  - Segment markets based on purchasing behavior.
  - Explain variation in consumer preferences.
  - Dempster et al. (1977): EM algorithm for latent data problems.
- The EM and Data Augmentation (DA, aka "Bayesian EM") algorithms exemplify the tight connection between LVM and efficient computation.

Machine Learning and Artificial Intelligence

- **Motivation**: Learn low-dimensional structure in high-dimensional data.
  - **Tipping & Bishop (1999)**: Probabilistic PCA.
- **Topic Modeling**:
  - Latent topics explain word distributions in documents.
  - Blei et al. (2003): Latent Dirichlet Allocation (LDA).
- **Autoencoders and VAEs**:
  - Learn latent representations for generative modeling.
  - Kingma and Welling (2014): Variational Autoencoders (VAEs).
  - **Mbacke, Clerc and Germain (2023)**: Statistical guarantees for VAEs via PAC Bayes.

Which LVM is used in this talk?

- The variable of interest W is sometimes impossible to measure directly (State of the economy, Traffic in a city, State of your health, State of a complex disease: Xu et al. (2016))
- Instead, one measures
  - $\mathbf{Y} = (Y_1, \ldots, Y_k)^T$ whose components are surrogates of W and each provide partial information about W
  - In addition, we measure also a covariate vector $\mathbf{X} \in \mathbb{R}^p$
- We are often interested in the explanatory power of $\mathbf{X}$ for W.

Copulas: The Joys

- Copulas are mathematical devices used to model dependence between random variables regardless of their marginals.
- Copulas are useful for data fusion/integration because they lead to coherent joint models, even when the marginals are in different families (e.g., Gaussian, Poisson, Student, etc) or of different types (e.g, discrete, continuous).
- Copulas unlock information contained in the dependence part of the distribution (second-order) that complements the information in the marginals.
- Simply put, copulas allow us to extend statistical methods beyond the use of a multivariate Gaussian or Student.

At the root of it all, a theorem

- If $Y_1, Y_2, \ldots, Y_K$ are continuous r.v.'s with cdfs $F_1, F_2, \ldots, F_k$, there is an unique copula $C : [0,1]^K \to [0,1]$ that links the joint cdf with the marginal ones (Sklar's Theorem).
- The copula (when $K = 2$) $C : [0,1] \times [0,1] \to [0,1]$ satisfies

$$F_{12}(t,s) = \Pr(Y_1 \leq t, Y_2 \leq s) = C(F_1(t), F_2(s)).$$

- The conditional copula satisfies

$$F_{12|X}(t,s) = \Pr(Y_1 \leq t, Y_2 \leq s|X) = C(F_{1|X}(t), F_{2|X}(s)|X)$$

At the root of it all, a theorem

- Usually we use parametric families so $C(u, v) = C_\theta(u, v)$ such as
  Clayton's family: $C_\theta(u, v) = \left[\max\left(u^{-\theta} + v^{-\theta} - 1, 0\right)\right]^{-1/\theta}$.
  Frank's family: $K_\theta(u, v) = -\frac{1}{\theta} \ln\left[1 + \frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right]$.

- In a conditional copula, θ may depend on X.

An example

- Cardiotocography (CTG) is a medical procedure that monitors the fetal heart rate.
- The LV is the fetus' underlying state of health during birth, W.
- Our surrogate response is the bivariate vector (Q, Y) where
  - Q is the number of peaks (acceleration followed by a deceleration of heart beats) for the signal recorded by the CTG
  - Y is the log of mean short-term "beat-to-beat" variability (MSTV) where the short-term variability (STV) is obtained by measuring the time between successive R waves (cardiac systoles) of the fetus' electrocardiogram.
- The covariates are `FM` (fetal movement) and `UC` (uterine contraction), two continuous variables monitored during birth.

Conditional independence LV model

- A canonical LV model, given $W_i = X_i\beta + \epsilon$, is

$$
\begin{aligned}
&Y_i \perp Q_i | W_i \\
&Y_i \sim N(\mu_c + \lambda_c W_i, \sigma^2) \\
&Q_i \sim Poisson(\exp(\mu_d + \lambda_d W_i))
\end{aligned}
$$

- This implies that the two marginal regressions share a common random effect so they are marginally dependent (and conditionally independent)
- The induced dependence is not analytically available.

Conditional independence is a Copula LV

- The copula alternative is, conditional on W_i,

$$H(Y_i, Q_i|W_i) = C_{\theta_i}(F_Y(Y_i|W_i), F_Q(Q_i|W_i)), \quad \theta_i = \kappa^{-1}(\xi_0 + \xi_1 W_i)$$
$$Y_i \sim N(\mu_c + \lambda_c W_i, \sigma^2); \; Q_i \sim Poisson(\exp(\mu_d + \lambda_d W_i))$$

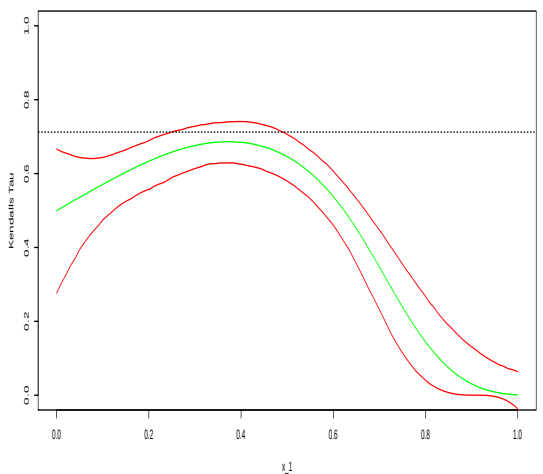
- The whole joint distribution of (Y, Q) is varying with W not just the marginals.
- The copula captures the residual dependence on W after the marginal effects have been accounted for.
- The previous model is obtained when the copula is the independence copula.
- When marginalizing over W we end up with a conditional copula model in which both marginals and the dependence structure vary with the covariates.

Why the Conditional Copula?

- $Y_i|x \sim N(f_i(x), \sigma_i)$ $x \in \mathbb{R}^2$
- True marginal means:
  - $f_1(x) = 0.6\sin(5x_1) - 0.9\sin(2x_2)$
  - $f_2(x) = 0.6\sin(3x_1 + 5x_2)$
  - $\sigma_1 = \sigma_2 = 0.2$, $X_1 \perp X_2$.
- Copula: $\theta(x) = 0.71$
- Suppose x_2 is not observed so inference is based only on x_1

Why the Conditional Copula?

- $Y_i|x \sim N(f_i(x), \sigma_i)$ $x \in \mathbb{R}^2$
- True marginal means:
  - $f_1(x) = 0.6\sin(5x_1) - 0.9\sin(2x_2)$
  - $f_2(x) = 0.6\sin(3x_1 + 5x_2)$
  - $\sigma_1 = \sigma_2 = 0.2$, $X_1 \perp X_2$.
- Copula: $\theta(x) = 0.71$
- Suppose x_2 is not observed so inference is based only on x_1



(Levi and Craiu, 2018)

CTG: The LV Copula Model

- $(Q_i, Y_i)|W_i$ has joint density

$$f_{(Q,Y)}(q,y) = f_c(y) \cdot \left[C_{d|c}\left(F_d(q), F_c(y)\right) - C_{d|c}\left(F_d(q-), F_c(y)\right) \right],$$

where

$$C_{d|c}(u_d, u_c) = \frac{\partial}{\partial u_c} C(u_d, u_c).$$

- Data Augmentation: Introduce latent variable Z such that

$$Q \stackrel{d}{=} F_d^-(F_Z(Z)),$$

- The copula between (Y, Z) is the same as the copula between (Y, Q)
- We can choose the distribution of Z to help the computation.
- For instance if we use a Gaussian copula, it helps to have $Z \sim N(0, 1)$
- Craiu and Sabeti (2012); Smith and Khaled (2012).

CTG: The Augmented LV Copula Model

- The dependence between Y, Z and Q is defined by their joint conditional distribution

$$f_{(Q,Z,Y)|W}(q,z,y \mid w) = h(z,y \mid w, \mu_c, \lambda_c, \psi_c, \boldsymbol{\xi}) \cdot \mathbb{1}_{F_Z^{-1}(F_d(q-|\varphi_d(\mu_d,\lambda_d,w)))\leq z<F_Z^{-1}(F_d(q|\varphi_d(\mu_d,\lambda_d,w)))}.$$

- Let $\boldsymbol{\xi} = (\xi_0, \xi_1) \in \mathbb{R}^2$ and $A(w) = \xi_0 + \xi_1 \cdot w$. Then we set

$$\theta(w, \boldsymbol{\xi}) = \frac{e^{A(w)} - e^{-A(w)}}{e^{A(w)} - e^{-A(w)}}$$

as the correlation parameter of the bivariate Gaussian conditional copula of $(Y, Z)|W = w$.

- Parameters are a priori independent

Some MCMC details

- If the copula and marginals are Gaussian the joint is a multivariate normal so some of the conditional densities are available in closed form.
- For other copula families we rely on MwG moves.
- We sample $\{Z_i : 1 \leq i \leq n\}$ from its conditional distribution and use the samples only to update the copula parameters $\boldsymbol{\xi}$.
- To update the remaining parameters, we condition on the observed data.
- Adaptive strategy for all MwG: target an acceptance rate of 44%.

Model Selection: WAIC

- The WAIC is defined as

$$\text{WAIC}(\mathcal{M}) = -2\text{fit}(\mathcal{M}) + 2\text{p}(\mathcal{M}), \tag{1}$$

where the model fitness is

$$\text{fit}(\mathcal{M}) = \sum_{i=1}^{n} \log\left(\mathbb{E}\left[\Pr(y_i, q_i|\omega, \mathcal{M})\right]\right) \tag{2}$$

and the penalty

$$\text{p}(\mathcal{M}) = \sum_{i=1}^{n} \text{Var}\left(\log\left(\Pr(y_i, q_i|\omega, \mathcal{M})\right)\right), \tag{3}$$

where ω contains all the parameters and latent variables in the model.

Spotlight on dependence: A conditional WAIC

- We use the following two conditional WAICs (Levi and Craiu, 2018)

$$\begin{aligned}\text{CWAIC}_{Y|Q}(\mathcal{M}) = &-2\sum_{i=1}^{n}\log\left(\mathbb{E}\left[\Pr(y_i|q_i,\omega,\mathcal{M})\right]\right)+\\ &+2\sum_{i=1}^{n}\text{Var}\left(\log\left(\Pr(y_i|q_i,\omega,\mathcal{M})\right)\right),\\ \text{CWAIC}_{Q|Y}(\mathcal{M}) = &-2\sum_{i=1}^{n}\log\left(\mathbb{E}\left[\Pr(q_i|y_i,\omega,\mathcal{M})\right]\right)+\\ &+2\sum_{i=1}^{n}\text{Var}\left(\log\left(\Pr(q_i|y_i,\omega,\mathcal{M})\right)\right),\end{aligned}$$

Spotlight on dependence: A conditional WAIC

- One can show (Levi and Craiu, 2018) $\text{CWAIC}_{Y|Q}$ is asymptotically equivalent to CCV for the marginal likelihood

$$\text{CCV}_{Y|Q}(\mathcal{M}) = \sum_{i=1}^{n} \log\left(\Pr(y_i|q_i, \mathcal{D}_{-i}, \mathcal{M})\right).$$

- Similarly, $\text{CWAIC}_{Q|Y}$ is asymptotically equivalent to

$$\text{CCV}_{Q|Y}(\mathcal{M}) = \sum_{i=1}^{n} \log\left(\Pr(q_i|y_i, \mathcal{D}_{-i}, \mathcal{M})\right).$$

Simulation Experiment

- Generate data using a Gaussian copula

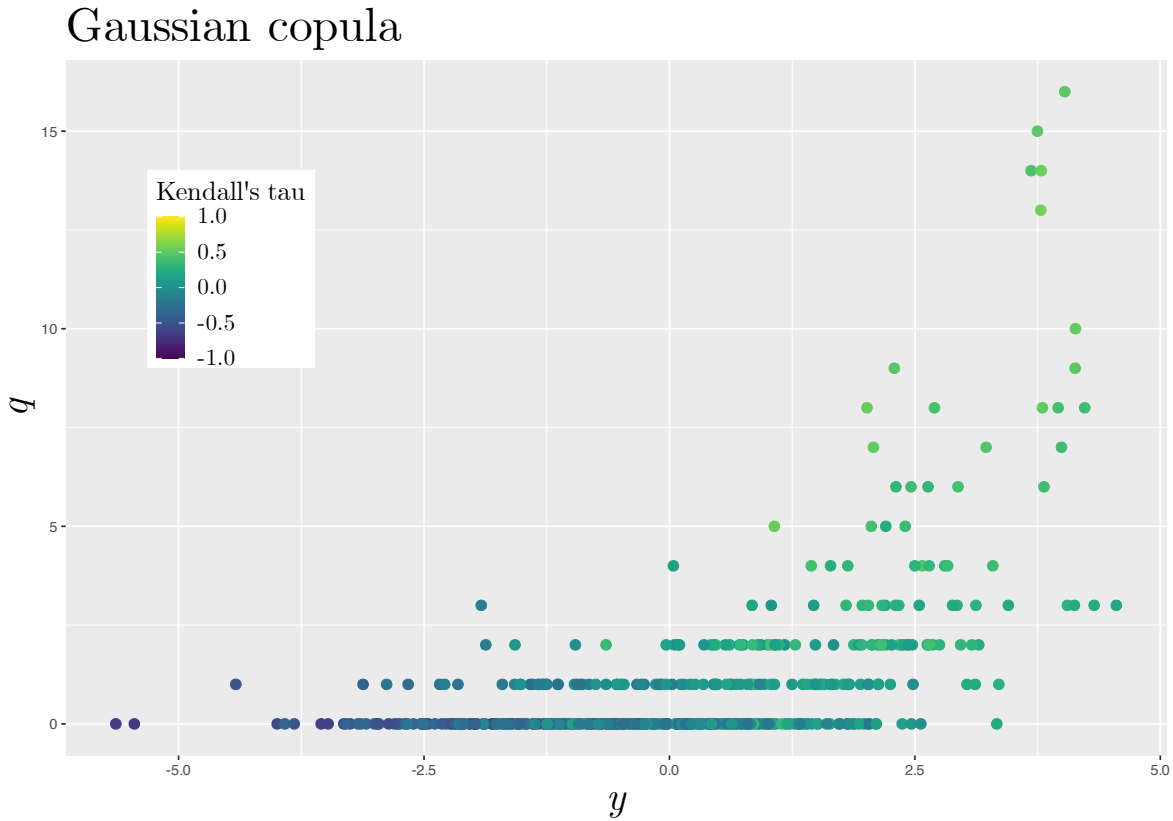


Figure: Bivariate scatterplot of the generated data with Gaussian copula, and Poisson and normal marginals

Simulation Experiment

- $\text{CWAIC}_{Y|Q}$ and $\text{CWAIC}_{Q|Y}$ selection criteria

Criteria\Copula	Gaussian	Frank	Gumbel	Clayton	Indep
$\text{CWAIC}_{Y\|Q}$	1627.36	1642.36	2395.17	1637.17	1606.31
$\text{CWAIC}_{Q\|Y}$	950.71	982.42	1673.57	976.05	997.43
Average	1289.04	1312.39	2034.37	1306.61	1301.87

Simulation Experiment

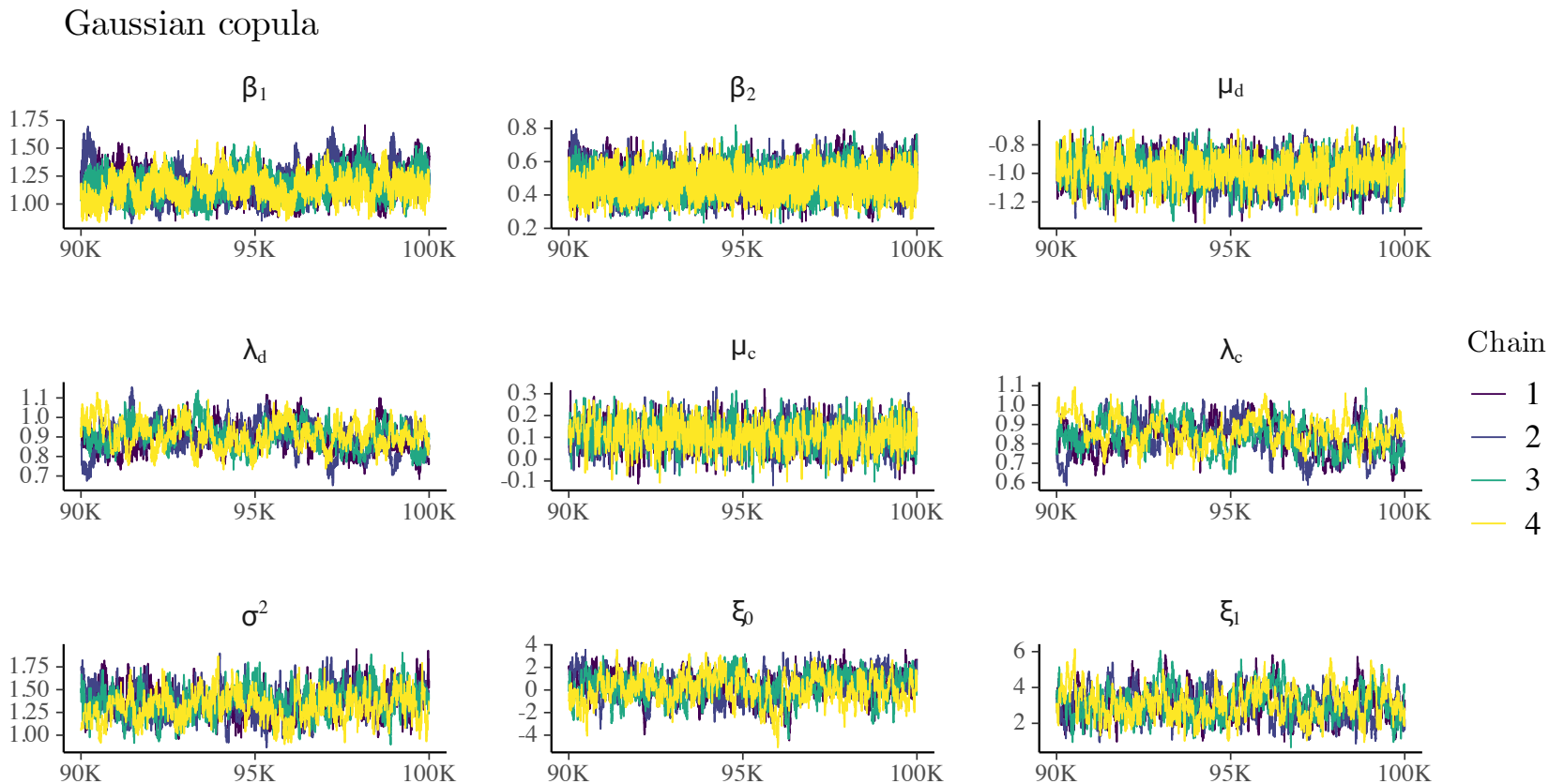
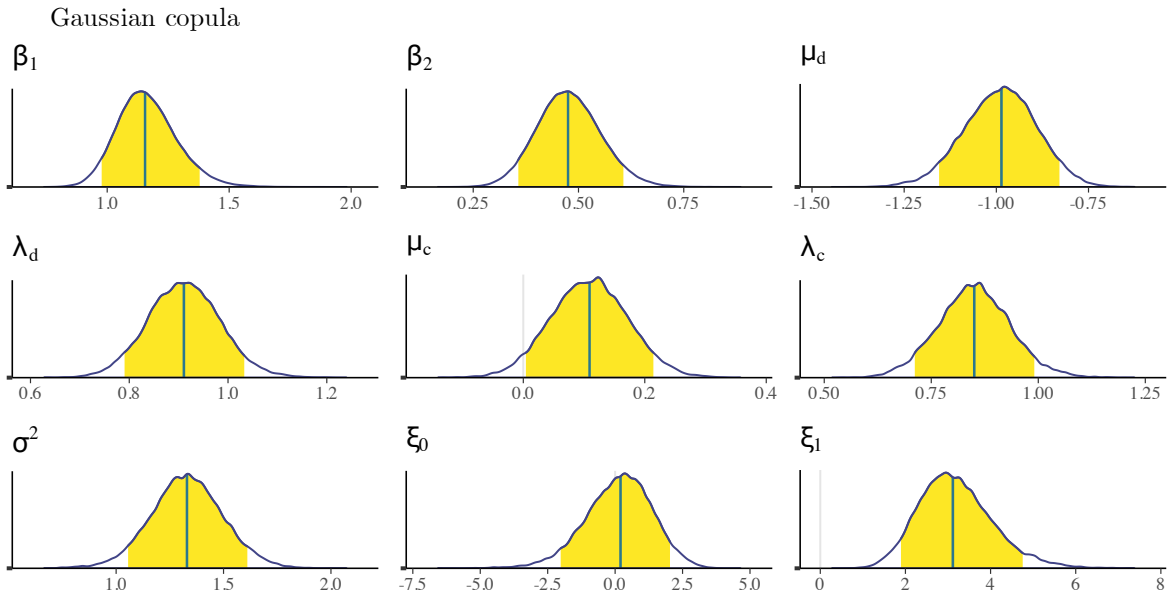


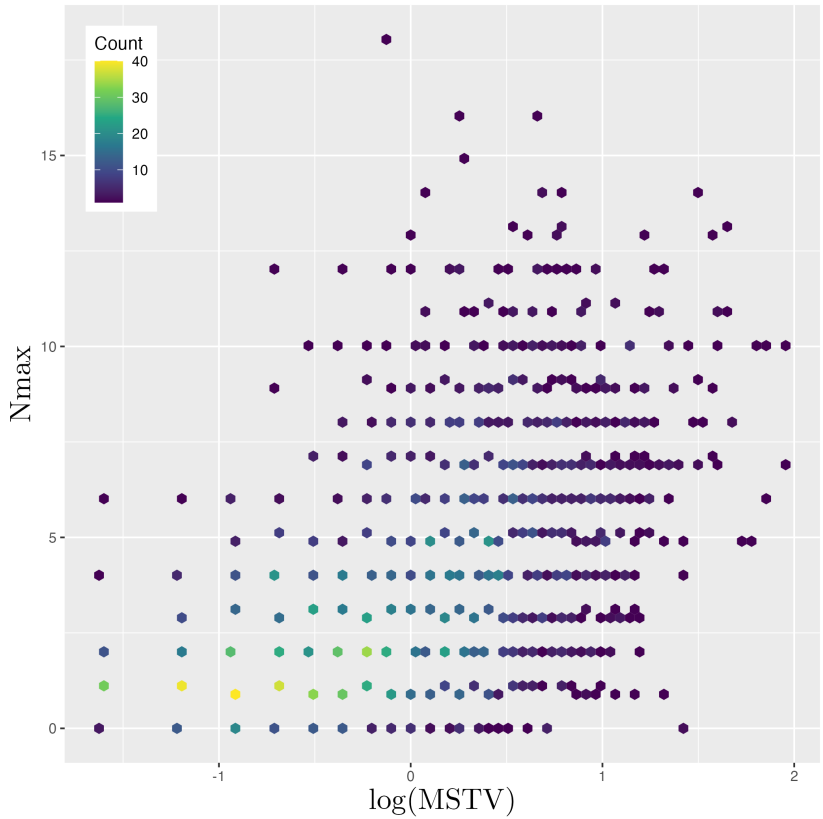
Figure: Traceplots for $\boldsymbol{\eta}$'s components.

Simulation Experiment



	β_1	β_2	λ_d	λ_c	ξ_1
Mean	1.18	0.48	0.90	0.84	3.10
True	1	0.5	1	1	3

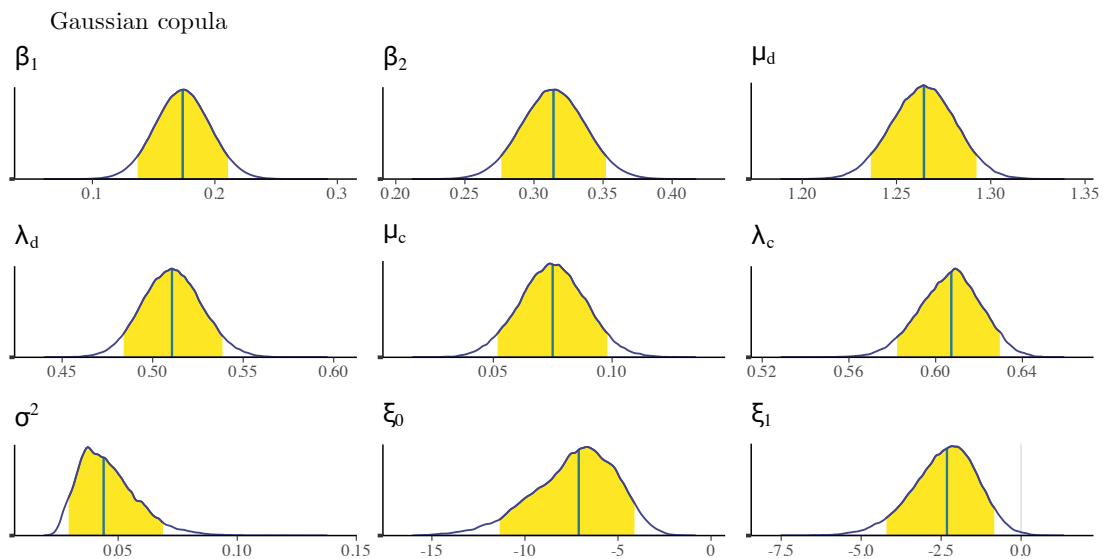
CTG: The data



CTG: Estimates

- *WAIC*, $WAIC_{Y|Q}$ and $WAIC_{Q|Y}$ all point to the Gaussian copula (over Gumbel, Frank, Clayton, Independence).
- The posterior means

	β_1 (FM)	β_2 (UC)	λ_d	λ_c	ξ_1
Mean	0.1744	0.3147	0.5101	0.6038	-2.3401



CTG: What does it mean?

- A peak in the histogram (counted with `Nmax`) would typically be produced by an FHR acceleration followed by a deceleration.
- Certain decelerations can be attributed to compression of the baby's head during uterine contractions, so they're not unusual.
- Late decelerations (starting after a uterine contraction begins) and especially variable decelerations often suggest a compromise in the supply of blood and oxygen to the fetus.
- A reduced STV can signify a quiet or sleep phase of the fetus, but also the effects of analgesic drugs given to the mother, fetal hypoxia, prematurity, neurological damage and tachycardia from any cause.
- Interpretation: Extremes values of W are identified with "unhealthy" regimes while small values of $|W|$ correspond to healthy ones.
- It is physiologically plausible that MSTV should be negatively correlated with Nmax.

References

Anderson, T. W. and Rubin, H. (1956). Statistical inference in factor analysis. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* **5** 111–150.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research* **3** 993–1022.

Craiu, R. V. and Sabeti, A. (2012). In mixed company: Bayesian inference for bivariate conditional copula models with discrete and continuous outcomes. *J. Multivariate Anal.* **110** 106–120.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B* **39** 1–38.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* **24** 417–441.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* .

Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent Structure Analysis*. Houghton Mifflin.

Levi, E. and Craiu, R. V. (2018). Bayesian inference for conditional copulas using Gaussian process single index models. *Computational Statistics & Data Analysis* **122** 115–134.

Smith, M. S. and Khaled, M. A. (2012). Estimation of copula models with discrete margins via bayesian data augmentation. *Journal of the American Statistical Association* **107** 290–303.

Spearman, C. (1904). General intelligence, objectively determined and measured. *The American Journal of Psychology* **15** 201–293.

Thurstone, L. L. (1935). *The Vectors of Mind*. University of Chicago Press.

Xu, L., Craiu, R. V., Sun, L. and Paterson, A. D. (2016). Parameter expanded algorithms for bayesian latent variable modeling of genetic pleiotropy data. *Journal of Computational and Graphical Statistics* **25** 405–425.

My papers are available on my homepage: `raducraiu.com`