Universidad Carlos III de Madrid

Variational Inference for high dimensional factor copulas

Hoang Nguyen

joint work with **M. Concepcion Ausin, Pedro Galeano**
Universidad Carlos III de Madrid

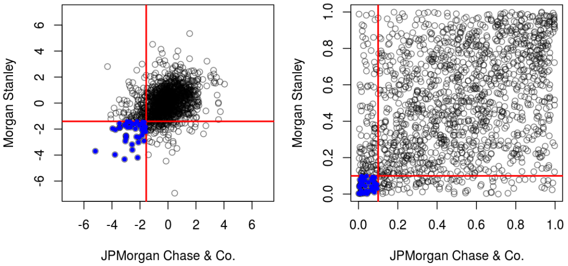
Tuesday 7th November, 2017

Introduction to Copulas

- A multivariate copula is a multivariate cdf defined on $[0,1]^d$ with uniform $U(0,1)$ marginals.
- Consider a n-dimensional joint cdf F with marginals $F_1, ..., F_d$. There exists a copula C, such that

$$F(x_1, ..., x_d) = C(F_1(x_1), ..., F_d(x_d))$$

for all x_i in $[-\infty, \infty]$, $i = 1, ..., d$.



Elliptical copulas

$$C_R^{Ga}(u_1,\dots,u_d) = \Phi_R^n(\Phi^{-1}(u_1),..,\Phi^{-1}(u_d))$$

$$C_{R,\nu}^{St}(u_1,\dots,u_d) = F_{R,\nu}^{MSt}(F_{t_\nu}^{-1}(u_1),..,F_{t_\nu}^{-1}(u_d))$$

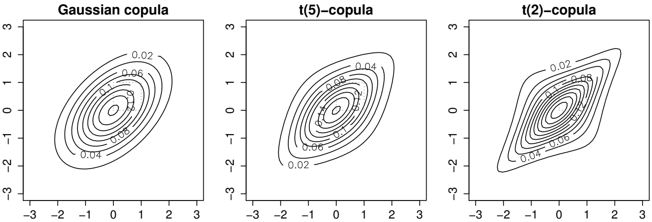


Figure: Contours of bivariate distributions with the same marginal standard normal

Archimedean copulas

Common Bivatiate Archimedean Copulas:

$$C(u_1, u_2) = \varphi^{-1}(\varphi(u_1) + \varphi(u_2))$$

Clayton (1978)	Frank (1979)	Gumbel (1960)
$\alpha \geqslant 0$	$\alpha \geqslant 0$	$\alpha \geqslant 1$
$\varphi(t) = t^{-\alpha} - 1$	$\varphi(t) = -ln\frac{e^{-\alpha t}-1}{e^{-\alpha}-1}$	$\varphi(t) = (-lnt)^{\alpha}$

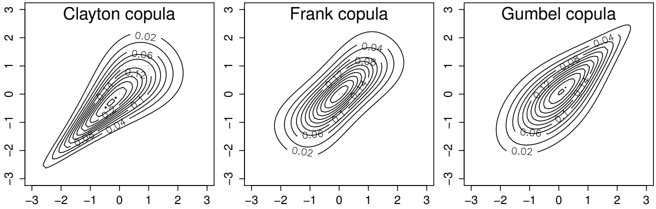


Figure: Contours of bivariate distributions with the same marginal standard normal

Vine copulas

Vine copula: C-vine, D-vine, R-vine (Aas et al., 2009)

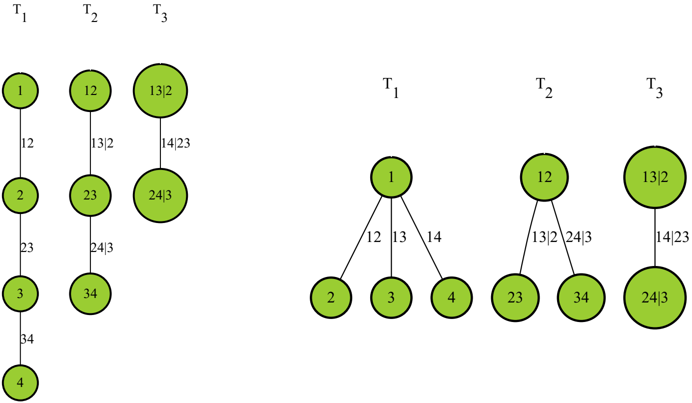


Figure: D-vine and Canonical vine copula

Factor copulas

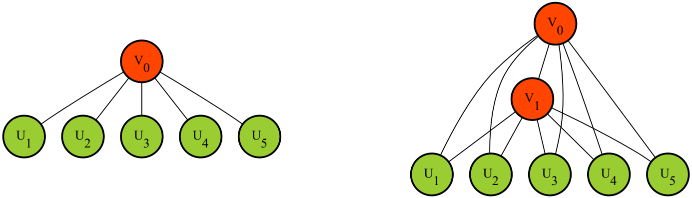


Figure: One factor and two factor copula models (Krupskii and Joe, 2013)

Bifactor and nested factor copulas

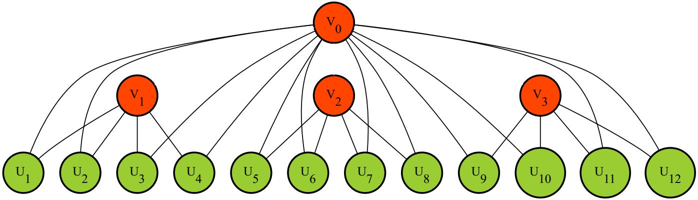


Figure: Bifactor copulas with $d = 12$ and $G = 3$ (Krupskii and Joe, 2015)

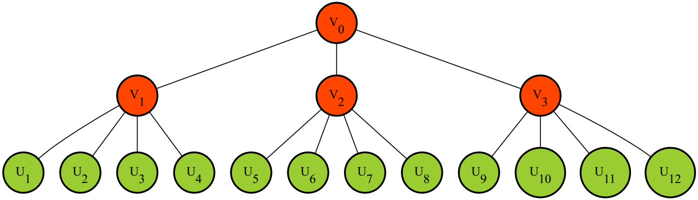


Figure: Nested factor copulas with $d = 12$ and $G = 3$ (Krupskii and Joe, 2015)

Posterior inference

Assuming that we have specify a factor copula structure together with bivariate linking copula in each tree layers.

- We are interested in the inference on the collection of latent variables and copula parameters $\{v, \theta\}$ based on the observables $\{u\}$
- The posterior is

$$p(v, \theta | u) = \frac{p(v, \theta, u)}{p(u)}$$

  - One factor copula, for example

$$\begin{aligned} p(v_0, \theta | u_1, \ldots, u_d) &\propto \prod_{i=1}^{d} \frac{p(u_i, v_0 | \theta)}{p(v_0)} p(v_0) p(\theta) \\ &\propto \prod_{i=1}^{d} c_{u_i, v_0}(u_i, v_0 | \theta) p(\theta) \end{aligned}$$

Posterior inference

For bifactor copula, we derive the posterior using the properties for vine copula,

$$p(v_0, v_1, \dots, v_G, \theta | u_1, \dots, u_d) \propto \prod_{g=1}^{G} \prod_{i=1}^{d_g} c(u_{i_g}, v_0, v_g | \theta) p(\theta)$$
$$\propto \prod_{g=1}^{G} \prod_{i=1}^{d_g} c_{u_{i_g}, v_0}(u_{i_g}, v_0 | \theta)$$
$$\times \prod_{g=1}^{G} \prod_{i=1}^{d_g} c_{u_{i_g}, v_g | v_0}(u_{i_g | v_0}, v_g | \theta) p(\theta)$$

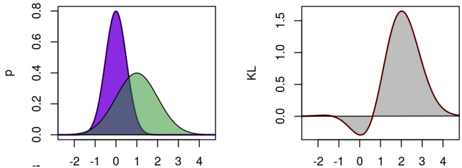
where $u_{i_g | v_0} = F(u_{i_g | v_0})$. Thus, it is computational expensive. We approximate the posterior by a proposal $q(v, \theta | \lambda^*)$.

$$q(v, \theta | \lambda^*) \approx p(v, \theta | u)$$

Kullback Leibler divergence

Variational Inference measures the different between two distributions using Kullback Leibler divergence:

$$KL(Q||P) = \int q(x) \log \frac{q(x)}{p(x)} dx \geq 0$$



Note that: $KL(Q||P) \neq KL(P||Q) \geq 0$

Objective function

We specify a family $\mathcal{Q}$ of densities as the proposal distribution

$$q(v,\theta|\lambda^*) = \arg\min_{\lambda} KL(q(v,\theta)||p(v,\theta|u))$$

$$KL(q(v,\theta)||p(v,\theta|u)) = \mathbb{E}_q[\log q(v,\theta)] - \mathbb{E}_q[\log p(v,\theta|u)]$$

$$KL(q(v,\theta)||p(v,\theta|u)) = \mathbb{E}_q[\log q(v,\theta)] - \mathbb{E}_q[\log p(v,\theta,u)] + \log p(u)$$

Objective function

We specify a family $\mathcal{Q}$ of densities as the proposal distribution

$$q(v,\theta|\lambda^*) = \arg\min_{\lambda} KL(q(v,\theta)||p(v,\theta|u))$$
$$KL(q(v,\theta)||p(v,\theta|u)) = \mathbb{E}_q[\texttt{log} q(v,\theta)] - \mathbb{E}_q[\texttt{log} p(v,\theta|u)]$$
$$KL(q(v,\theta)||p(v,\theta|u)) = \mathbb{E}_q[\texttt{log} q(v,\theta)] - \mathbb{E}_q[\texttt{log} p(v,\theta,u)] + \texttt{log} p(u)$$

Because we cannot compute the KL, we optimize an alternative objective (Evidence lower bound) that is equivalent to the KL up to an added constant:

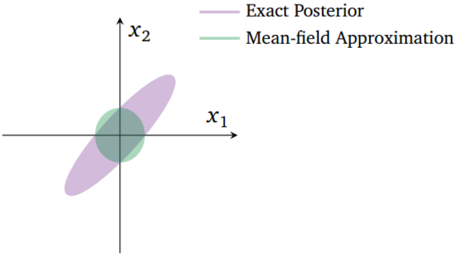
$$\begin{aligned}\texttt{ELBO}(q) &= \mathbb{E}_q[\texttt{log} p(v,\theta,u)] - \mathbb{E}_q[\texttt{log} q(v,\theta)] \\ &= \texttt{log} p(u) - KL(q(v,\theta)||p(v,\theta|u)) \leq \texttt{log} p(u)\end{aligned}$$

when $q(v,\theta) = p(v,\theta|u)$, we obtain $\texttt{ELBO} = \texttt{log} p(u)$

Mean field variational family

In mean-field variational family, the latent variables are mutually independent and each governed by a distinct factor in the variational density.

$$q(v,\theta) = \prod_{l=1}^{\#latents} q(v_l) \prod_{i=1}^{\#\theta} q(\theta_i)$$



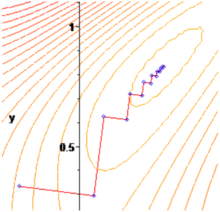
Black Box Variational Inference

We specify a family $\mathcal{Q}$ of densities over the latent variables.

$$\lambda^* = \arg\max_{\lambda} \mathbb{E}_{q(v,\theta)}[\log p(v,\theta,u)] - \mathbb{E}_{q(v,\theta)}[\log q(v,\theta)]$$

such that $supp(q(v,\theta|\lambda)) \subseteq supp(p(v,\theta|u))$

- We could propose directly a density approximation $q(v,\theta|\lambda)$ and take the derivative wrt. λ
- Update $\lambda = \lambda + Step * Gradient$



- However, this direct approach produces noisy evaluations of the gradient, $\nabla_\lambda \left(\mathbb{E}_{q(v,\theta)}[\log p(v,\theta,u)] - \mathbb{E}_{q(v,\theta)}[\log q(v,\theta)]\right)$.

Black box variational inference

An automated algorithm (ADVI) to solve the optimization problem based on continuous transformations of the parameters (Kucukelbir, 2016).

- Define a one-to-one differentiable function.

$$T : supp(p(v, \theta | u)) \longrightarrow \mathbb{R}^K$$

- Any continuous transformation could be possible:
  - Correlation constrain: $T(\theta) = \texttt{atanh}\theta = \frac{1}{2}\log\left(\frac{1+\theta}{1-\theta}\right)$
  - Positive constrain: $T(\theta) = \texttt{log}(\theta)$
  - Lower constrain: $T(\theta) = \texttt{log}(\theta - L)$
  - Lower and upper bound constrain: $T(\theta) = \texttt{logit}\frac{\theta - L}{U - L}$

Variance reduction technique

The optimization becomes:

$$\mu^*, \sigma^* = \arg\max_{\mu,\sigma} \mathbb{E}_{N(\mu,\sigma)}[\log p(v, \theta, u)] - \mathbb{E}_{N(\mu,\sigma)}[\log q(v, \theta)]$$

- Draw M samples $\eta \sim \mathcal{N}(0, I)$.
- Obtain $x_k = \mu_k + \eta_k \sigma_k$.
- Obtain $(v_k, \theta_k) = T^{-1}(x_k)$
- Average over M samples for the ELBO.
- Similar approach to calculate the gradient of ELBO. Update μ, σ
- This algorithm is guaranteed to converge to a local maximum of the ELBO under certain conditions on the step-size sequence.
- Because $\sigma > 0$, we optimize over $\omega = \log \sigma$ instead

Automatic Differentiation Variational Inference in Stan

Algorithm 1: Automatic differentiation variational inference

Data: Copula Data $U = \{u_i\}$
Result: The value μ, ω
Initialization $\mu^{(0)} = 0, \omega^{(0)} = 0$;
while *Any change in copula types* **do**
  while *Change in* ELBO *is above some threshold* **do**
    Draw M samples $\eta_m \sim N(0, 1)$;
    Invert the standardized $x_m = \mu^{(j)} + \exp(\omega^{(j)})\eta_m$;
    Approximate the noisy gradient $\nabla_\mu \text{ELBO}$ and $\nabla_\omega \text{ELBO}$;
    Update $\mu^{(i+1)} \leftarrow \mu^{(j)} + \varrho^{(j)} \nabla_\mu \mathcal{F}$;
    Update $\omega^{(i+1)} \leftarrow \omega^{(j)} + \varrho^{(j)} \nabla_\omega \mathcal{F}$;
    Incremental iteration (i) ;
  end
  Select best bivariate copula u_i and v based on AIC,BIC ;
  Reassign the copulas and estimate ;
end
Return Copula structure and the parameters of proposal distribution;

One factor copula model

We generate a sample of d = 100 variables with T = 1000 time observations. Bivariate copula types are Gaussian, Student, Clayton, Gumbel, Frank, Joe (and their rotation 90, 180, 270 degree) and Mix copulas. Time is report in seconds using one core Intel i7-4770 processor.

Table: Time of Computation and Copula selection

Copula type	Gaussian	Student	Clayton	Gumbel	Frank	Joe	Mix
Initial at correct structure							
Time estimated (s)	6	322	18	24	5	9	59
ELBO	31181	35490	78769	67530	58375	76254	58438
Initial at random structure							
Time estimated (s)	303	625	325	258	316	308	382
Selection iteration	3	3	4	2	3	4	4
% correction	98%	78%	62%	100%	100%	57%	88%
ELBO	31191	35410	78767	67539	58383	76277	58449

(about 100 - 200 paramters / 100 bivariate copulas / 1 latent factor)

One factor copula model

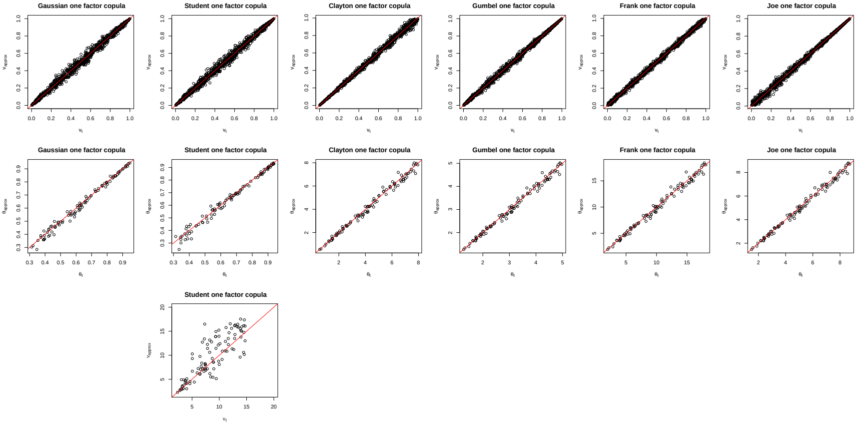


Figure: Posterior means of v and θ versus true values

One factor copula model

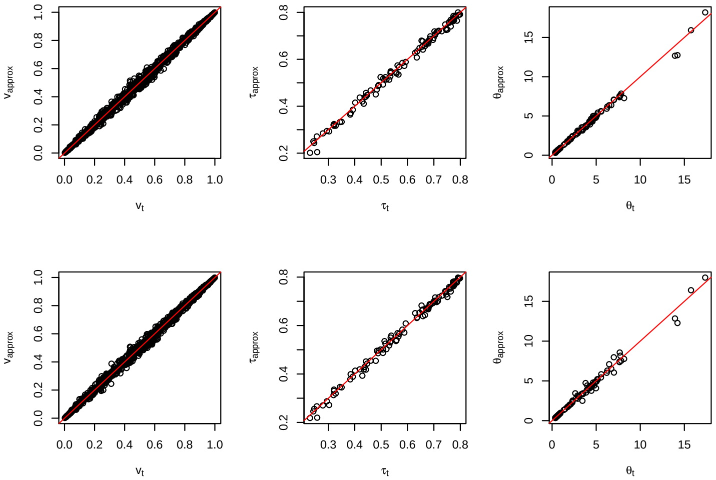


Figure: Mixed copula estimation with a correct vs random initial structure

Nested factor copula model

We generate the nested factor copula with d = 100 variables, N = 1000 observations and G = 5 groups of latent factors.

Table: Time of Computation and Copula selection

Copula type	Gaussian	Student	Clayton	Gumbel	Frank	Joe	Mix
Initial at correct structure							
Time estimated (s)	7	334	18	27	9	11	80
ELBO	24731	25351	69358	59615	47988	69989	41796
Initial at random structure							
Time estimated (s)	379	1045	354	417	333	380	481
Selection iteration	4	5	5	5	4	6	5
% correction	72%	72%	70%	97%	97%	58%	79%
ELBO	22595	25209	68966	57550	46157	69993	41804

(about 105 - 210 paramters / 105 bivariate copulas / 6 latent factors)

Nested factor copula model

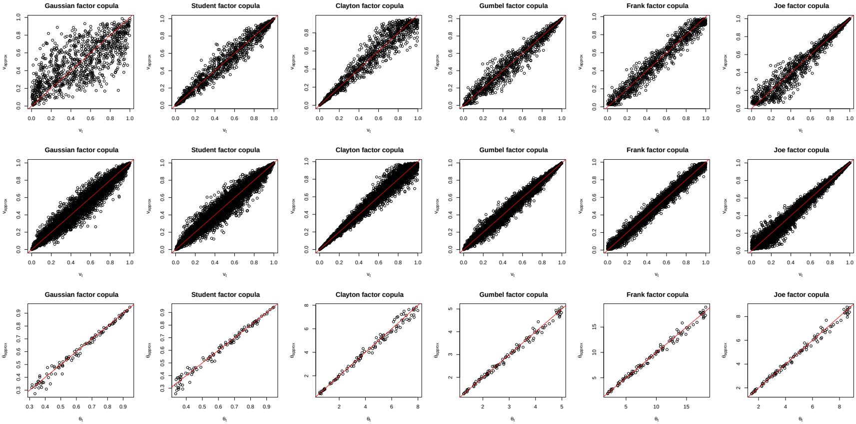


Figure: Posterior means of v_0, v_g and θ versus true values

Bifactor copula model

We generate the bifactor copula with d = 100 variables, T = 1000 time observations and G = 5 groups of latent factors.

Table: Time of Computation and Copula selection

Copula type	Gaussian	Student	Clayton	Gumbel	Frank	Joe	Mix
Initial at correct structure							
Time estimated (s)	59	1212	119	102	56	100	515
ELBO	50413	83977	136734	117332	96655	135002	93867
Initial at random structure							
Time estimated (s)	1589	4317	857	1028	743	718	1025
Selection iteration	4	6	6	4	6	5	6
% correction Tree 1	99%	69%	82%	100%	99%	48%	77%
% correction Tree 2	97%	79%	76%	57%	98%	44%	66%
ELBO	51260	83917	136508	111419	99575	134895	96287

(about 200 - 300 paramters / 200 bivariate copulas / 6 latent factors)

Bifactor copula model

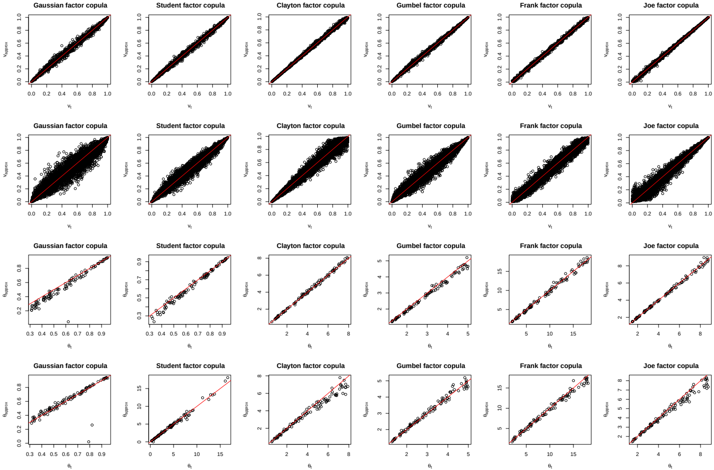


Figure: Posterior means of v_0, v_g and θ versus true values

Financial return dependence

We illustrate an empirical example using d = 100 stock returns divided into G = 10 groups from 01/01/2010 to 31/12/2013 of the companies listed in S&P 500 index. The daily data contain T = 1000 observation days. We use AR(1)-GARCH(1,1)to marginalize each stock returns:

$$
\begin{aligned}
r_{it} &= c_i + \phi_{i1} r_{i,t-1} + a_{it} \\
a_{it} &= \sigma_{it} \eta_{it} \\
\sigma^2_{it} &= \omega_i + \alpha_{i1} a^2_{i,t-1} + \beta_{i1} \sigma^2_{i,t-1}
\end{aligned}
$$

with skewed Student-t innovation, η_{it}. Then, the dependence structure of innovations is modelled by a factor copula function

$$
\begin{aligned}
\eta_{1t}, \ldots, \eta_{dt} &\sim F(\eta_{1t}, \ldots, \eta_{dt}) \\
&\sim C(F(\eta_{1t}), \ldots, F(\eta_{dt}) | \theta, v)
\end{aligned}
$$

Financial return dependence

Table: Time of Computation and Copula selection

Structure	One factor	Nested factor	Two factor	Bifactor copula
Time estimated (s)	1559	2225	4812	5059
ELBO	33340	34232	35051	36070
Selection iteration	3	5	6	4
# bivariate links	100	110	200	200
% Gaussian	0	4	1	12
% Student	94	90	71	92
% Clayton (rotated)	0	0	0	1
% Gumbel (rotated)	6	16	29	13
% Frank (rotated)	0	0	95	61
% Joe (rotated)	0	0	0	1
% Independence	0	0	3	12

Conclusion

- Fast variational inference for factor copula model in high dimentions.
- Copula bivariate selection based on VI estimation performs well with simulation data.
- Compared to MCMC, variational inference tends to be faster and easier to scale to large data.
- VI generally underestimates the variance of the posterior density. However, the relative accuracy of variational inference and MCMC is still unknown. But we obtain quite reasonable result with factor copula models with limited time.

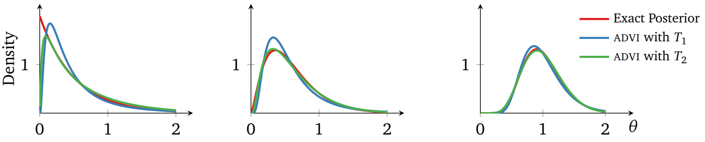
Sensitivity to Transformations

Consider a posterior density in the Gamma family, with support over $\mathbb{R}_{>0}$. Figure 9 shows three configurations of the Gamma, ranging from Gamma$(1, 2)$, which places most of its mass close to $\theta = 0$, to Gamma$(10, 10)$, which is centered at $\theta = 1$. Consider two transformations T_1 and T_2

$$T_1 : \theta \mapsto \log(\theta) \quad \text{and} \quad T_2 : \theta \mapsto \log(\exp(\theta) - 1),$$

both of which map $\mathbb{R}_{>0}$ to $\mathbb{R}$. ADVI can use either transformation to approximate the Gamma posterior. Which one is better?

Figure 9 show the ADVI approximation under both transformations. Table 2 reports the corresponding KL divergences. Both graphical and numerical results prefer T_2 over T_1. A quick analysis corroborates this. T_1 is the logarithm, which flattens out for large values. However, T_2 is almost linear for large values of θ. Since both the Gamma (the posterior) and the Gaussian (the ADVI approximation) densities are light-tailed, T_2 is the preferable transformation.



(a) Gamma$(1, 2)$ **(b)** Gamma$(2.5, 4.2)$ **(c)** Gamma$(10, 10)$

Figure 9: ADVI approximations to Gamma densities under two different transformations.