Expectation Propagation for the Estimation of Conditional Bivariate Copulas

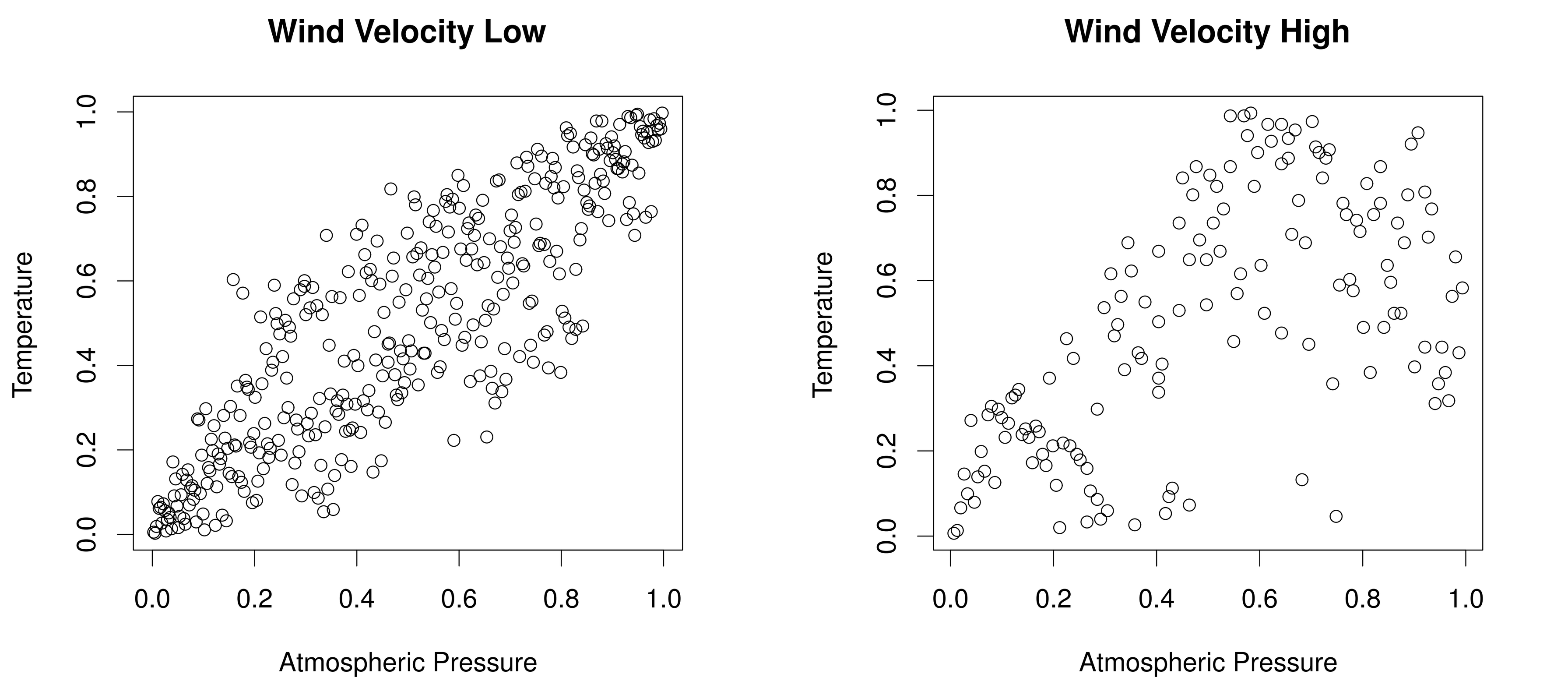
José Miguel Hernández-Lobato(1), David López-Paz(2) and Zoubin Ghahramani(1)

(1) - Department of Engineering, Cambridge University
(2) - Max Planck Institute for Intelligent Systems, Tübingen, Germany

Neural Information Processing Systems Foundation

Copulas and Covariates

A bivariate copula $\boldsymbol{C}$ may depend on a **covariate** denoted by $\boldsymbol{Z}$.



Conditional Copulas

We can adjust for the effect of $\boldsymbol{Z}$ by modelling the conditional copula.

Conditional Copula: [Patton 2006]
Let $F_{X|Z}$ and $F_{Y|Z}$ be the conditional marginals of X and Y given Z. The conditional copula C_Z for X and Y is the distribution of $F_{X|Z}(X|Z)$ and $F_{Z|Y}(Y|Z)$.

Extension of Sklar's Theorem: [Patton 2006]
Let $F_{X|Z}$ and $F_{Y|Z}$ be the conditional marginals of X and Y given Z. Let F_Z be the conditional distribution with continuous marginals $F_{X|Z}$ and $F_{Y|Z}$. Then, there is a unique conditional copula C_Z such that $F_Z(x, y|z) = C_Z[F_{X|Z}(x|z), F_{Y|Z}(y|z)|z]$.

The estimation of F_Z from a sample $\{X_i, Y_i, Z_i\}_{i=1}^n$ can then be done by

1. Estimating $F_{X|Z}$ and $F_{Y|Z}$ using our favorite method.
2. $\{X_i, Y_i\}_{i=1}^n$ to $[0,1]^2$ using the estimates for $F_{X|Z}$ and $F_{Y|Z}$.
3. Estimating C_Z using the data from the step 2. **We need a model for C_Z**.

A Semi-parametric Model for Conditional Copulas

We describe C_Z using a **parametric model** specified in terms of Kendal's tau.

Kendall's tau: [Joe 1997]
Let (U_1, V_1) and (U_2, V_2) be two independent samples from a bivariate copula $\boldsymbol{C}$.
Kendall's tau is $\tau = \mathcal{P}[(U_1 - U_2)(V_1 - V_2) > 0] - \mathcal{P}[(U_1 - U_2)(V_1 - V_2) < 0]$.

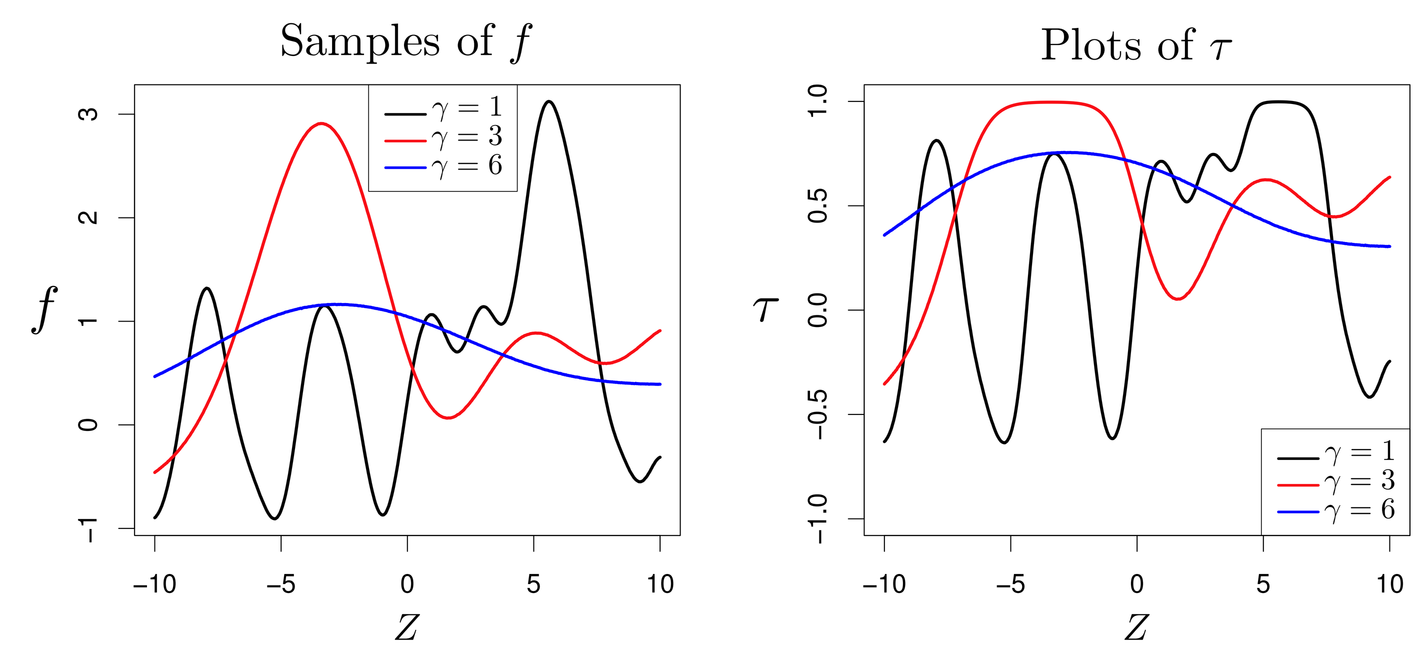
- ⋆ $\tau = 0 \leftrightarrow U$ and V are independent.
- ⋆ $\tau = 1 \leftrightarrow U$ and V have perfect positive dependence.
- ⋆ $\tau = -1 \leftrightarrow U$ and V have perfect perfect negative dependence.

Most parametric bivariate copulas are **fully determined given** τ.
For example, in the Gaussian copula, $\rho = \sin(\tau\pi/2)$.

The dependence of C_Z on Z is captured by the relationship $\tau = \sigma[f(Z)]$, where f is an arbitrary non-linear function and $\sigma(x) = 2\Phi(x) - 1$ is a sigmoid function.

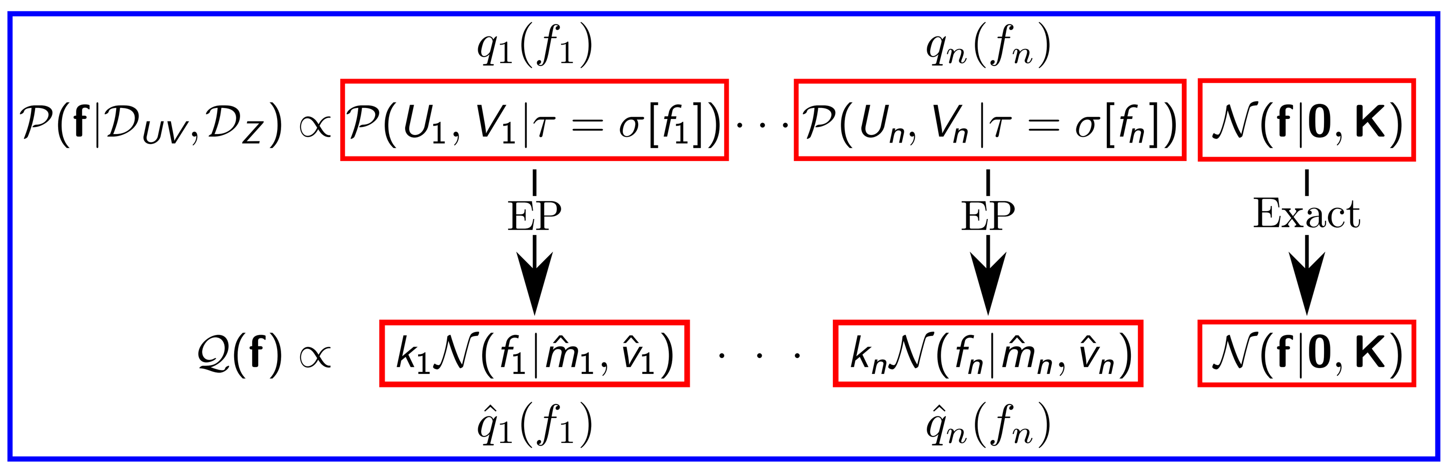
Bayesian Inference

We assume that f follows *a priori* a **Gaussian process** with zero mean and covariance function $k[f(Z_i), f(Z_j)] = \exp\{-0.5\gamma^{-2}(Z_i - Z_j)^2\}$.



Expectation Propagation [Minka 2001]

EP approximates the posterior $\mathcal{P}(\mathbf{f}|\mathcal{D}_{UV}, \mathcal{D}_Z)$ by a simpler distribution $\mathcal{Q}(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\mathbf{m}, \mathbf{V})$, where



EP tunes $\hat{m}_i$ and $\hat{v}_i$ by minimizing $\mathrm{KL}[q_i(f_i)\mathcal{Q}(\mathbf{f})[\hat{q}_i(f_i)]^{-1}||\mathcal{Q}(\mathbf{f})]$, $i = 1, \ldots, n$.
Similar to EP for Gaussian process classification [Rassmussen and Williams 2006].

We fix γ by maximizing the EP approx. of $\mathcal{P}(\mathcal{D}_{UV}|\mathcal{D}_Z)$. The total cost is $\mathcal{O}(n^3)$.

Experiments with Synthetic and Real-world Data

We choose a **Gaussian copula** for the likelihood function $\mathcal{P}(u, v|\tau = \sigma[f(z)])$.

Synthetic Data:
1000 points. Z sampled uniformly from $[0, \pi]$.
U and V sampled from a Gaussian copula with $\tau = 0.5\cos(3z)$.

Meteorological Data:
522 points. Atmospheric pressure (X), Temperature (Y) and Wind Velocity (Z).
Conditional marginals $F_{X|Z}$ and $F_{Y|Z}$ estimated using kernels [Hall et al. 2004].
These estimates are used to map X and Y into U and V.

Experimental Protocol:
Training sets with 100 observations.
40 repetitions. Test log-likelihood (**TLL**) used as a measure of performance.

Benchmark Methods:
⋆ **GC:** Parametric Gaussian copula with no dependence on Z.
⋆ **MLL:** Non-parametric maximum local likelihood estimator of f [Acar et al. 2011].

Results

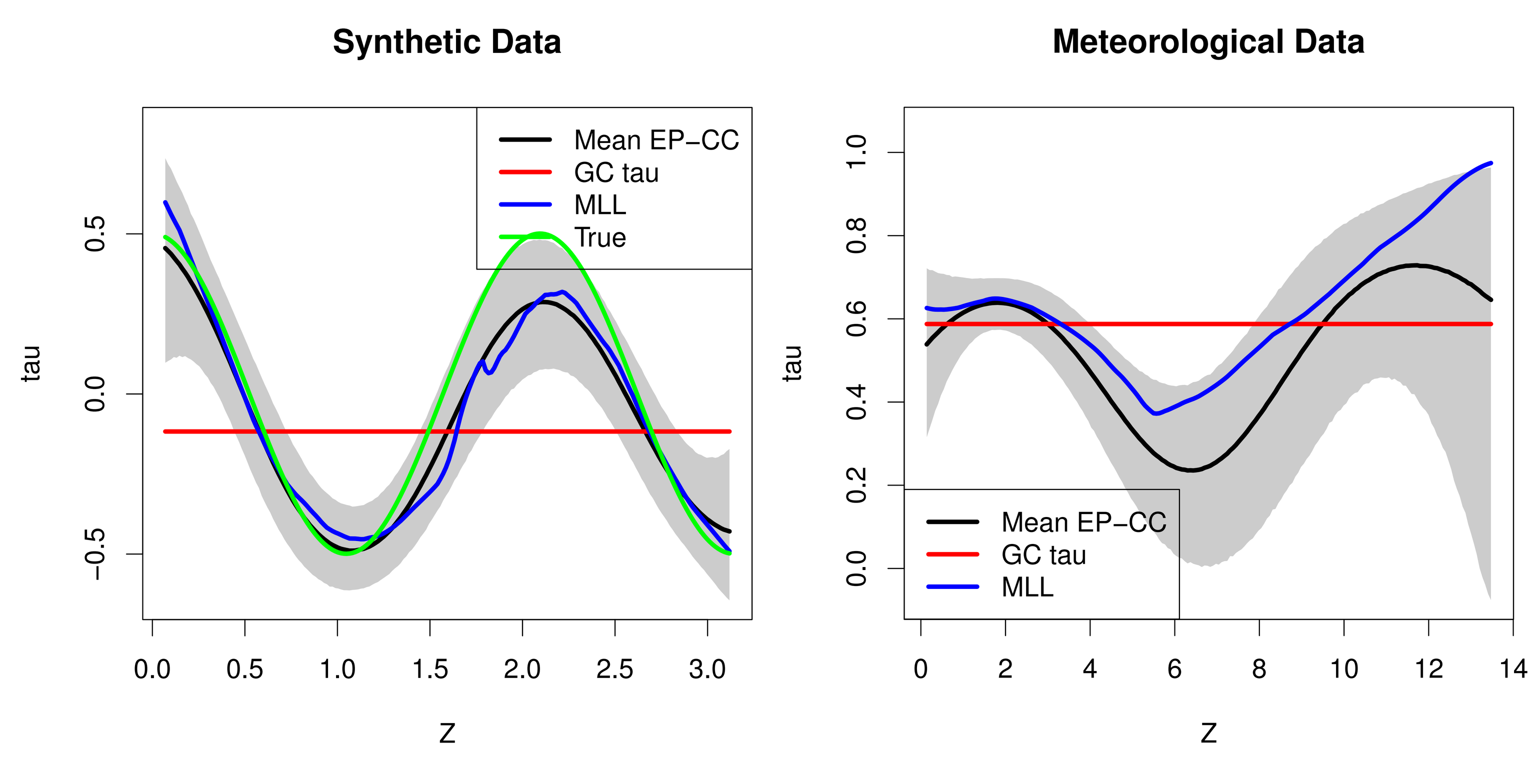
EP-CC: The EP method for the estimation of conditional copulas.

Dataset	Method	Avg. TLL	p-value[1]
Synthetic	**EP-CC**	**0.13±0.021**	-
	MLL	0.06±0.150	**0.002**
	GC	-0.01±0.017	$< \mathbf{2.2 \cdot 10^{-16}}$
Meteorological	**EP-CC**	**0.49± 0.019**	-
	MLL	0.42±0.479	**0.35**
	GC	0.47±0.029	$\mathbf{7.03 \cdot 10^{-5}}$

1: p-value of a paired Student's t test between EP-CC and the other methods.

EP-CC is **more robust** than MLL.
The meteorological data have a copula that **depends** on the wind velocity.

Plots of f for each Dataset (200 training points)



Conclusions

- We have proposed a **semi-parametric** model for conditional copulas.
  The model employs a **Gaussian process** prior.
  Expectation propagation is used for efficient Bayesian inference.
- The performance of this method (EP-CC) has been illustrated in experiments with **synthetic** and **meteorological** data.
- In these experiments, EP-CC is more **robust** and can perform better than other approaches that maximize the local likelihood.

References

1. A. J. Patton. Modelling Asymmetric Exchange Rate Dependence, International Economic Review, 2006.
2. H. Joe. Multivariate models and dependence concepts. Chapman & Hall, 1997.
3. T. Minka. A family of algorithms for approximate Bayesian inference. MIT, 2001.
4. C. E. Rasmussen and C. K. I. Williams. Gaussian Processes for Machine Learning. MIT Press, 2006.
5. P. Hall, J. S. Racine and Q. Li. Cross-Validation and the Estimation of Conditional Probability Densities, Journal of the American Statistical Association, 2004, 99, 1015-1026.
6. E. F. Acar, R. V. Craiu and F. Yao. Dependence Calibration in Conditional Copulas: A Nonparametric Approach, Biometrics, 2011, 67, 445-453.