High-dimensional Copula Constructions in Machine Learning

Gal Elidan

The Hebrew University

What is this talk about?

Goal: Provide an overview of high-dimensional copula-based constructions in ML

I will:

- Describe the key components of several general purpose models
- Present sample results for each work
- Discuss central merits and relation to other works

Scope

- Learning with tree-averaged distributions [Kirshner, 2008]
- The Nonparanormal [Liu, Laffery, Wasserman, JMLR 2009]
- Copula Bayesian Networks [Elidan, NIPS 2010]
- Copula Processes [Wilson and Ghahramani, NIPS 2010]

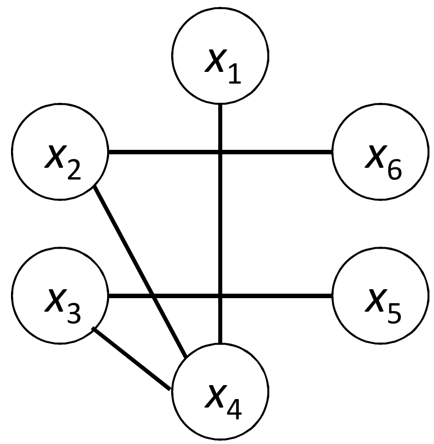
What will not be covered:

- Ricardo Silva's work (later today)
- Copula-based applications (a few are here today)
- Works that use copulas but do not directly aim to model joint distributions (we will also see some of those)
- Related constructions (some very interesting!) (e.g. cumulative distribution networks, Huang and Frey, 2008)

Markov Networks

U is an undirected graph that encodes independencies:

$$X_i \perp \mathcal{X} - \{X_i\} - N(X_i) | N(X_i)$$



where $N(X_i)$ are the neighbors of X_i in U

Theorem (Hammersley-Clifford):

If f is positive and the independencies hold then it factorizes according to U

For trees:

$$f_{\mathcal{X}}(\mathbf{x}) = \left[\prod_i f_i(x_i)\right] \left[\prod_{(i,j)\in E} \frac{f_{i,j}(x_i, x_j)}{f_i(x_i) f_j(x_j)}\right]$$

Bayesian mixture of all trees

Challenge: there are $N^{(N-2)}$ trees

Idea: use edge weight matrix β to define a prior over trees

$$P(T \mid \beta) = \frac{1}{Z} \prod_{(i,j)\in T} \beta_{i,j} \quad \text{with} \quad Z = \sum_{T} \prod_{(i,j)\in T} \beta_{i,j}$$

Theorem (Meila and Jaakkla 2006):

1. Easy to compute Z (via generalized Laplacian matrix)
2. Decomposability of the prior allows us to compute average over all tree efficiently

$\Rightarrow$

Average density over copula trees (still a copula!) can be computed via ratio of matrix determinants

Kirshner, 2008

From Bivariate Copulas to Copula Trees

It follows that the joint copula also decomposes:

$$c_{\mathcal{X}}(\mathbf{x}) = \frac{f_{\mathcal{X}}(\mathbf{x})}{\prod_i f_i(x_i)}$$

Kirshner, 2008

From Bivariate Copulas to Copula Trees

It follows that the joint copula also decomposes:

$$c_{\mathcal{X}}(\mathbf{x}) = \frac{f_{\mathcal{X}}(\mathbf{x})}{\prod_i f_i(x_i)} = \prod_{(i,j)\in E} \frac{f_{i,j}(x_i, x_j)}{f_i(x_i) f_j(x_j)}$$

Kirshner, 2008

From Bivariate Copulas to Copula Trees

It follows that the joint copula also decomposes:

$$c_{\mathcal{X}}(\mathbf{x}) = \frac{f_{\mathcal{X}}(\mathbf{x})}{\prod_i f_i(x_i)} = \prod_{(i,j)\in E} \frac{f_{i,j}(x_i, x_j)}{f_i(x_i) f_j(x_j)} = \prod_{(i,j)\in E} c_{i,j}(x_i, x_j)$$

Given marginals, we can find the optimal tree efficiently using a maximum spanning tree algorithms

Upside: only bivariate estimation (different than vines!)

Downside: assumptions are too simplistic

Kirshner, 2008

Estimation using EM

Parameters: 1) the edge weight matrix β
2) the bivariate copula parameters θ_{ij}

E-Step: need to compute posterior over $N^{(N-2)}$ trees!

Decomposability $\Rightarrow$ need only compute $N(N-1)/2$ edge probabilities and reuse computations.

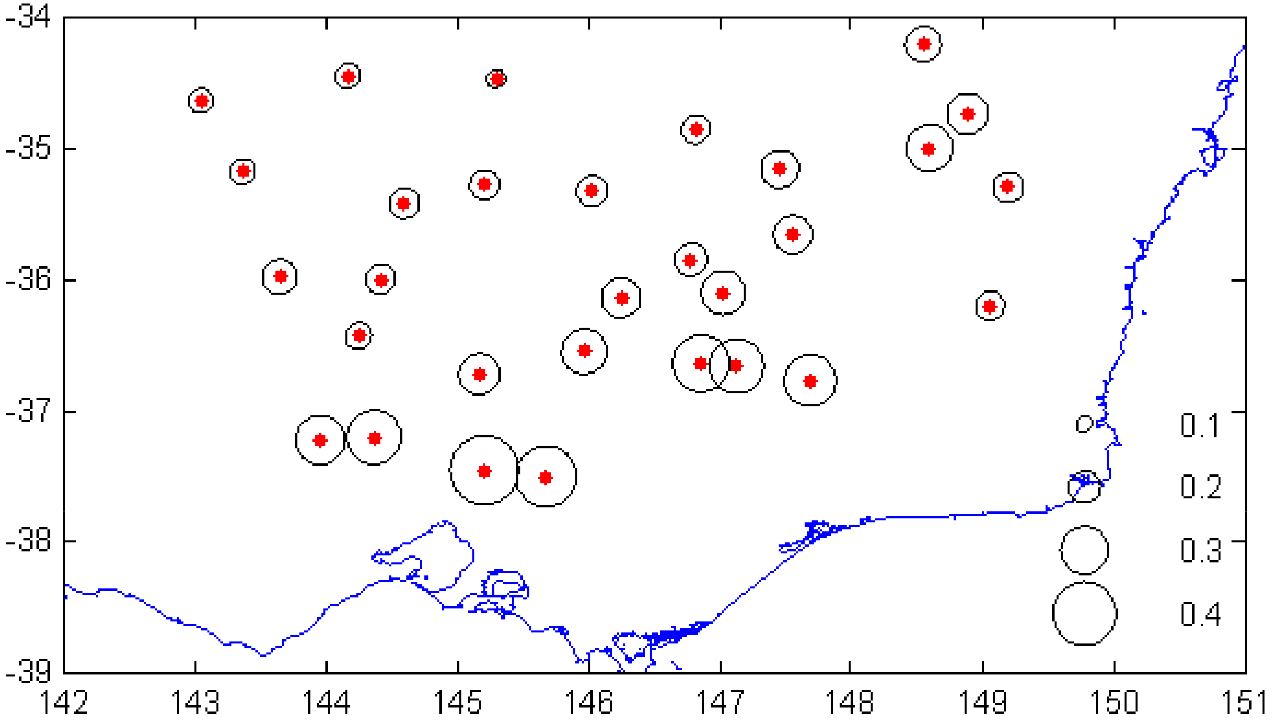
M-Step: standard optimization of bivariate copulas that depends only on pairs of variables

Assuming copula estimation complexity of $O(M)$: complexity of learning the model is $O(MN^3)$

Practical for tens of variables!

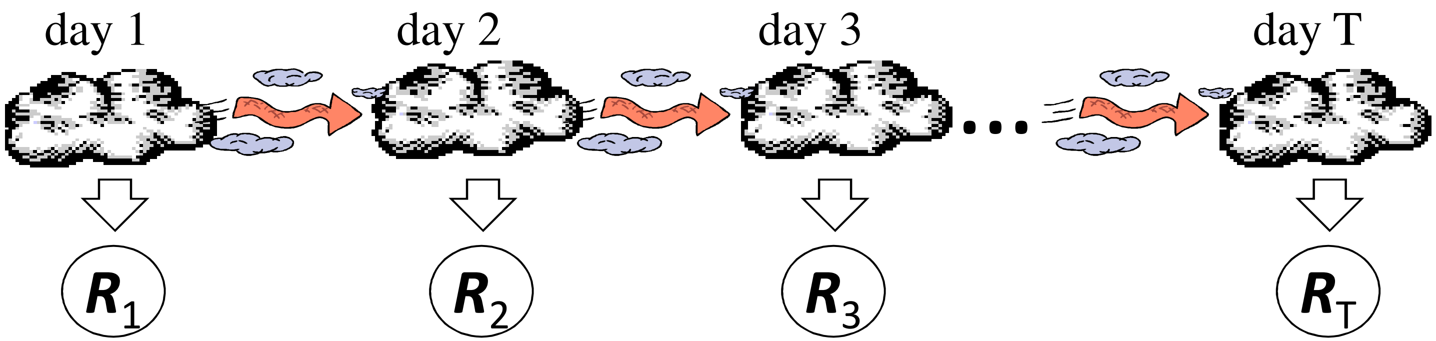
Kirshner, 2008

Modeling Daily Multi-Site Rainfall



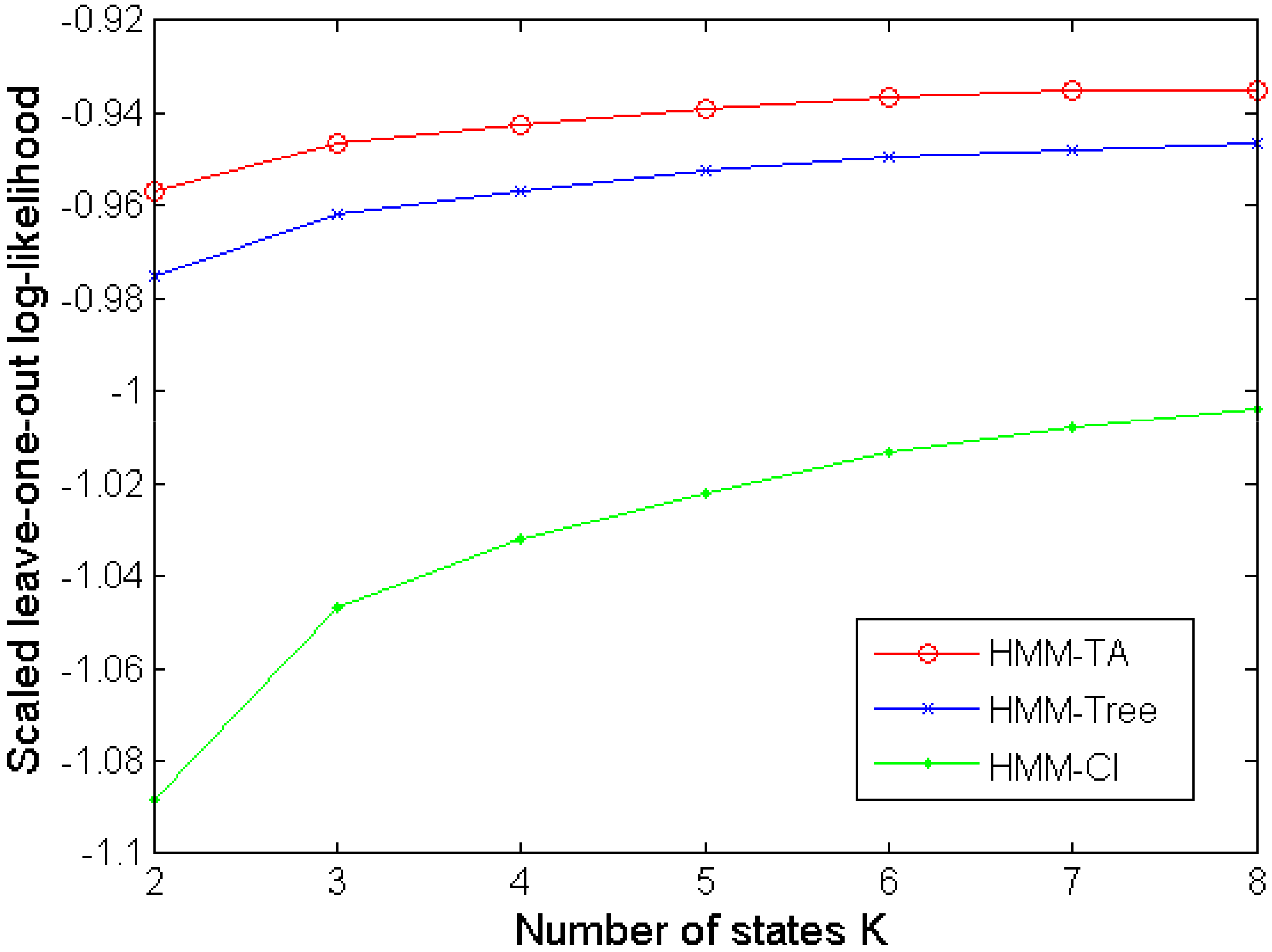
d stations
(10-40)

N observed days
(3000-8000)



Kirshner, 2008

Selecting Number of States



Kirshner, 2008

Consistent estimation in high-dimension

Assumptions	Dimension	Regression	Graphical Models
Parametric	Low	Linear model	Multivariate normal
	High	LASSO	Graphical LASSO
Nonparametric	Low	Additive model	?
	High	Sparse additive model	

Goal: theoretically founded estimation for nonparametric high-dimensional undirected graphs

Liu, Lafferty, Wasserman, 2009

The Nonparanormal Distribution

$X = (X_1,\ldots,X_p)^T \sim$ **NPN(μ,Σ,f)** if there exists univariate functions $\{f_j(X_i)\}$ such that

$$(f_1(X_1), \ldots, f_p(X_p)) \sim N(\mu, \Sigma)$$

Isn't this is just a Gaussian copula?

Yes, if $f_i(X_i)$ are monotone and differentiable

So what is the problem?

- High-dimensionality leads to estimation issues (p>n)
- Plugging in the empirical distribution does not work in the semiparametric case…

Liu, Lafferty, Wasserman, 2009

Density-less Structure Estimation

Let $h_j(x) = \Phi^{-1}(F_j(x))$ and Λ be the covariance of h(x)

Key insight: ($X_j \perp X_i$ | rest) if and only if $\Lambda_{ij}^{-1}=0$

can estimate structure solely from ranks

1. Replace observation with normal score

$$\tilde{f}_j(x) = \Phi^{-1}(\tilde{F}_j(x))$$

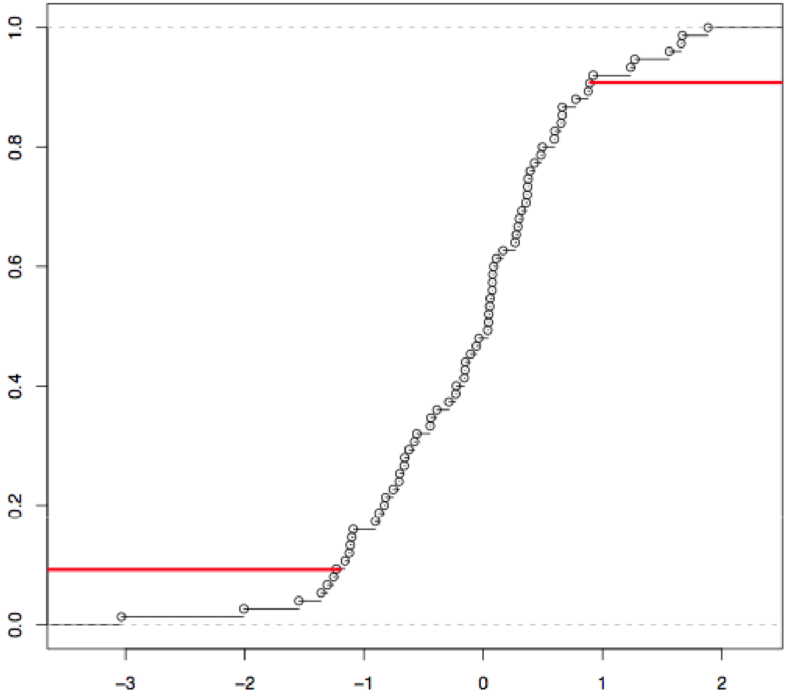
2. Compute functional sample covariance

$$S_n(\tilde{f}) = \frac{1}{n}\sum_{i=1}^{n} \tilde{f}(X[i])\tilde{f}(X[i])^T$$

3. Estimate structure from $S_n(\tilde{f})$ (e.g. using glasso)

Liu, Lafferty, Wasserman, 2009

Winsorized Estimator $\tilde{F}_j$



$\delta_n \equiv \frac{1}{4n^{1/4}\sqrt{\pi \log n}}$

$\delta_n \equiv \frac{1}{4n^{1/4}\sqrt{\pi \log n}}$

Main result: $\max_{i,j} \left| S_n(\tilde{f})_{ij} - S_n(f)_{ij} \right| = o_P(n^{-1/4})$

⇨

risk, norm (of Σ) and model selection consistency

(using analysis of Rothman et al, 2008, and Ravikumar, 2009)

Liu, Lafferty, Wasserman, 2009

Synthetic Structure Recovery

- 40 nodes
- 2 different transforms
- several training sample sizes

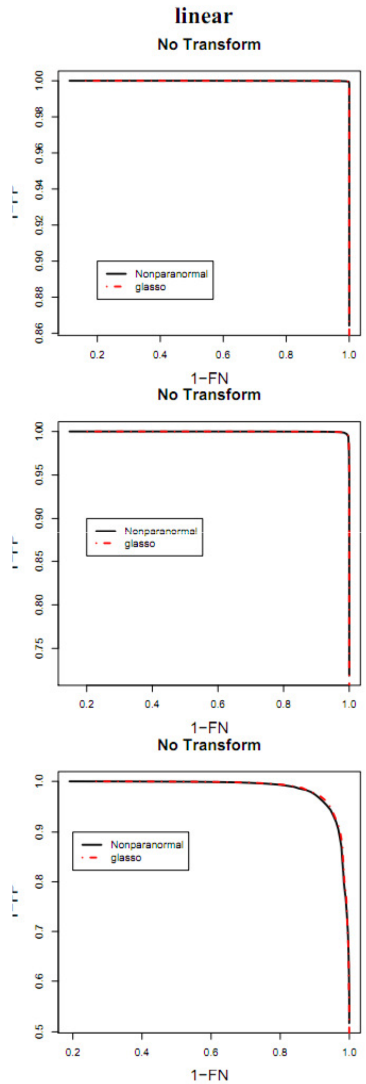


Figure 7: ROC curves for sample sizes $n = 1000, 500, 200$ (top, middle, bottom).

Liu, Lafferty, Wasserman, 2009

Synthetic Structure Recovery

- 40 nodes
- 2 different transforms
- several training sample sizes

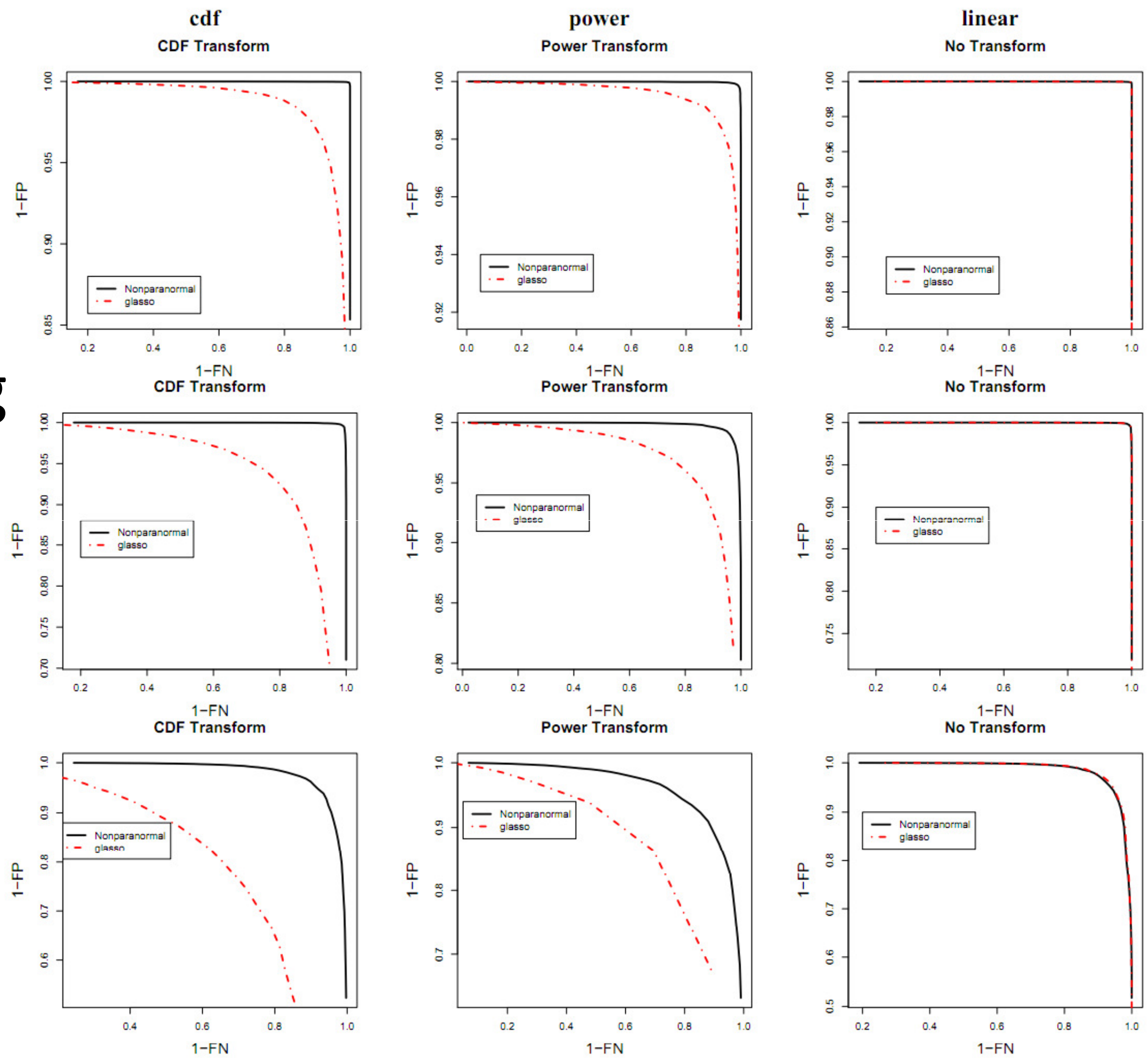
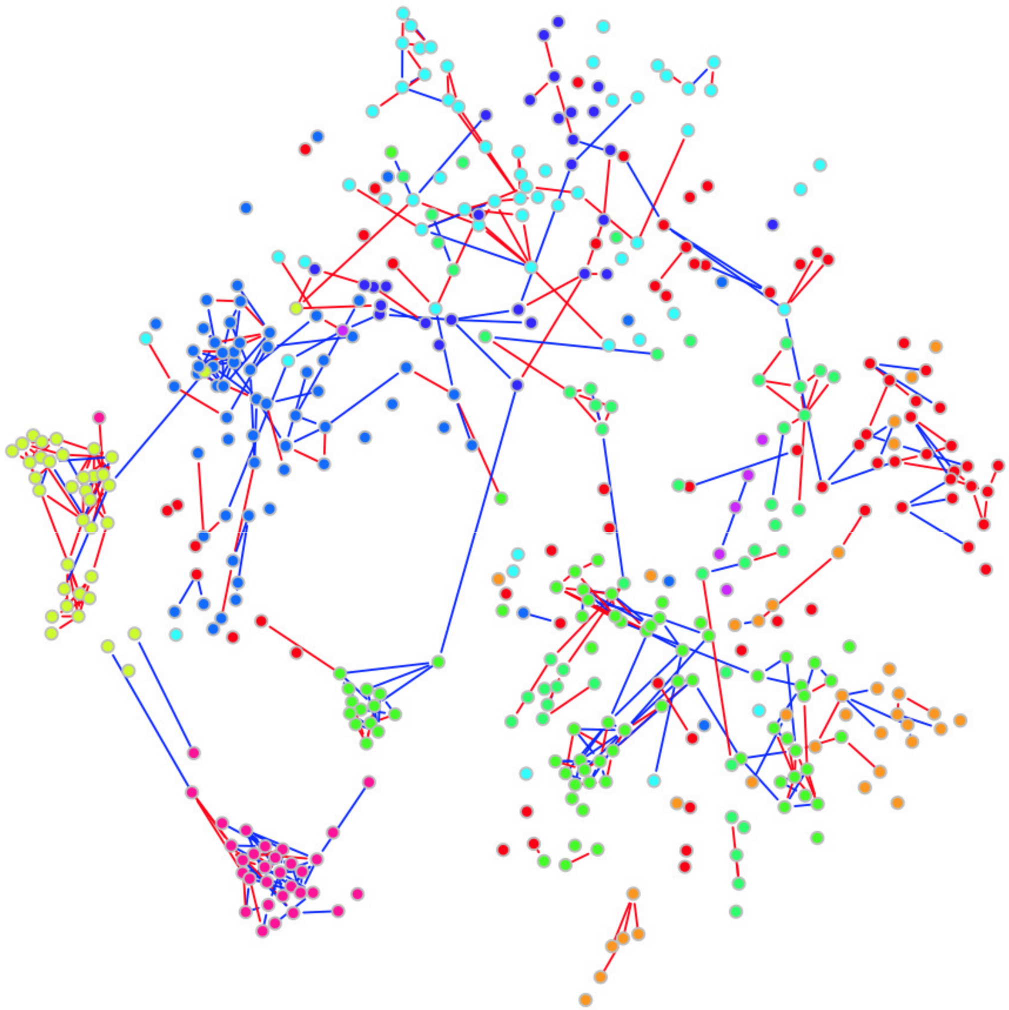


Figure 7: ROC curves for sample sizes $n = 1000, 500, 200$ (top, middle, bottom).

Liu, Lafferty, Wasserman, 2009

S&P 500: differences from glasso



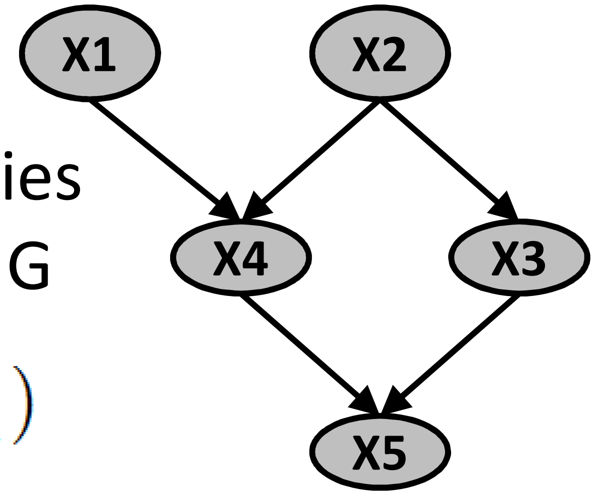
Non-Gaussian case possibly reveals new useful information

Liu, Lafferty, Wasserman, 2009

Bayesian Networks

G is a <u>directed</u> graph that encodes independencies:

$$X_i \perp \text{Non-descendents}_i | \text{Parents}_i$$



Theorem:

If f is positive and the independencies hold then it factorizes according to G

$$f_{\mathcal{X}}(\mathbf{x}) = \prod_i f_{i|par_i}(x_i \mid x_{par_i})$$

✓ Intuitive representation of uncertainty

✓ Easy to construct using local $f_{i|par_i}(x_i \mid x_{par_i})$

Conditional Densities Using Copulas

Simple bivariate case:

$$f(x|y) = \frac{f(x,y)}{f(y)}$$

(this is Kirshner presented differently)

Elidan, 2010

Conditional Densities Using Copulas

Simple bivariate case:

$$f(x|y) = \frac{f(x,y)}{f(y)} = \frac{c(F(x),F(y))f(x)f(y)}{f(y)}$$

(this is Kirshner presented differently)

Elidan, 2010

Conditional Densities Using Copulas

Simple bivariate case:

$$f(x|y) = \frac{f(x,y)}{f(y)} = \frac{c(F(x),F(y))f(x)f(y)}{f(y)} = c(F(x),F(y))f(x)$$

(this is Kirshner presented differently)

Theorem: For **any** f(x|**y**), there **exists** a copula such that

$$f(x|\mathbf{y}) \quad = \quad R_c(F(x), F(y_1), \ldots, F(y_K))f(x)$$

Elidan, 2010

Conditional Densities Using Copulas

Simple bivariate case:

$$f(x|y) = \frac{f(x,y)}{f(y)} = \frac{c(F(x),F(y))f(x)f(y)}{f(y)} = c(F(x),F(y))f(x)$$

(this is Kirshner presented differently)

Theorem: For **any** f(x|**y**), there **exists** a copula such that

$$\begin{aligned} f(x|\mathbf{y}) &= R_c(F(x), F(y_1), \ldots, F(y_K))f(x) \\ &\equiv \frac{c(F(x), F(y_1), \ldots, F(y_K))}{\frac{\partial^K C(1,F(y_1),\ldots,F(y_k))}{\partial F(y_1)\ldots\partial F(y_K)}} f(x) \end{aligned}$$

simpler than the copula density!

And constructive converse also holds!

Elidan, 2010

From local to global Copulas

Theorem: If the independencies in G hold then

$$c(F(x_1),\ldots,F(x_N)) = \prod_i R_{c_i}(F(x_i),\{F(\mathbf{pa}_{ik})\})$$

(and vice-versa)

→

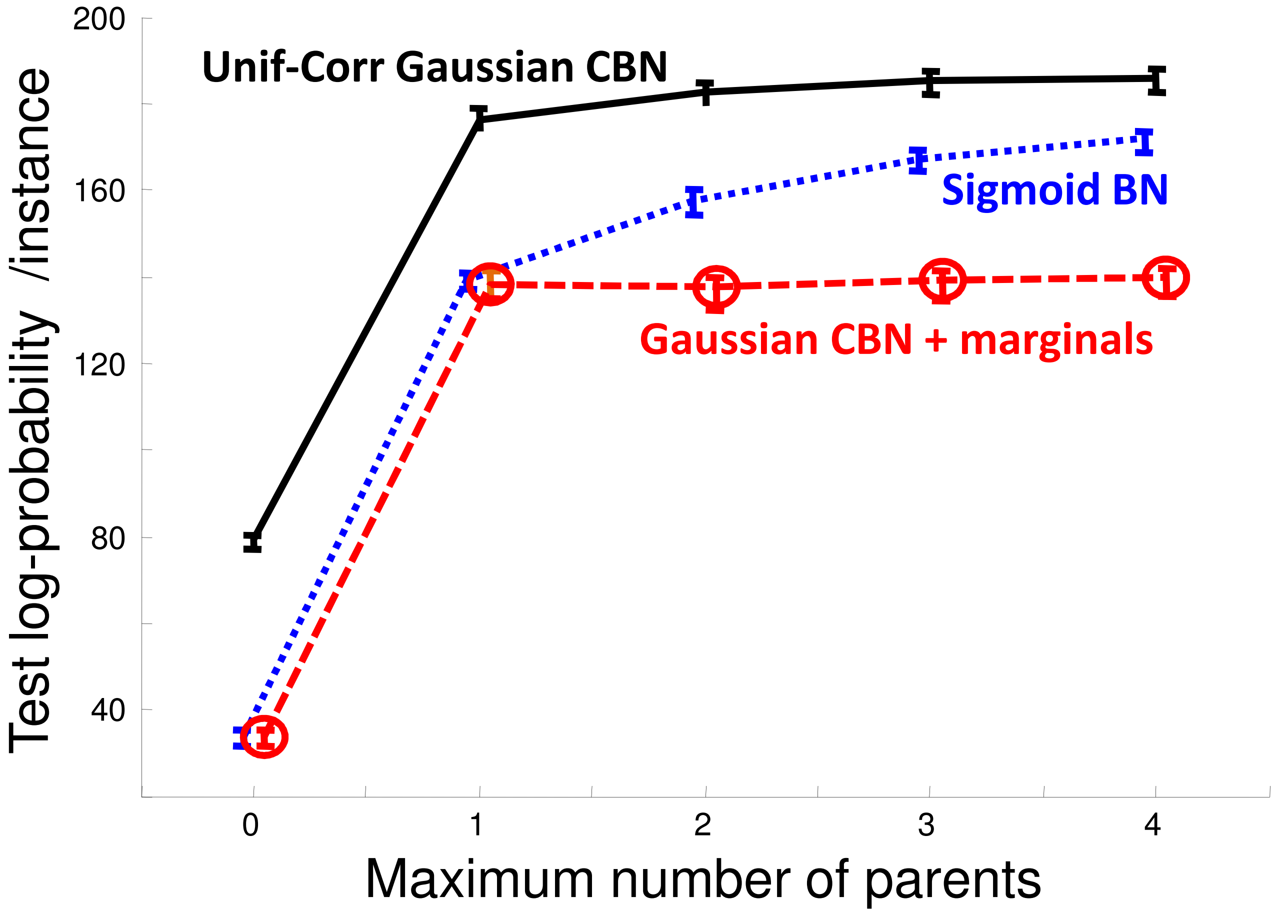
A **Copula Network** defines a valid joint density

$$f(\mathbf{x}) = \prod_i R_i(F(x_i),F(\mathbf{par}_{i1}),\ldots,F(\mathbf{par}_{ik_i}))f(x_i)$$

- Can now use standard estimation and graphical models structure learning techniques
- Similar to NPBBN (Hanea 2008), but avoids conditional rank correlations

Elidan, 2010

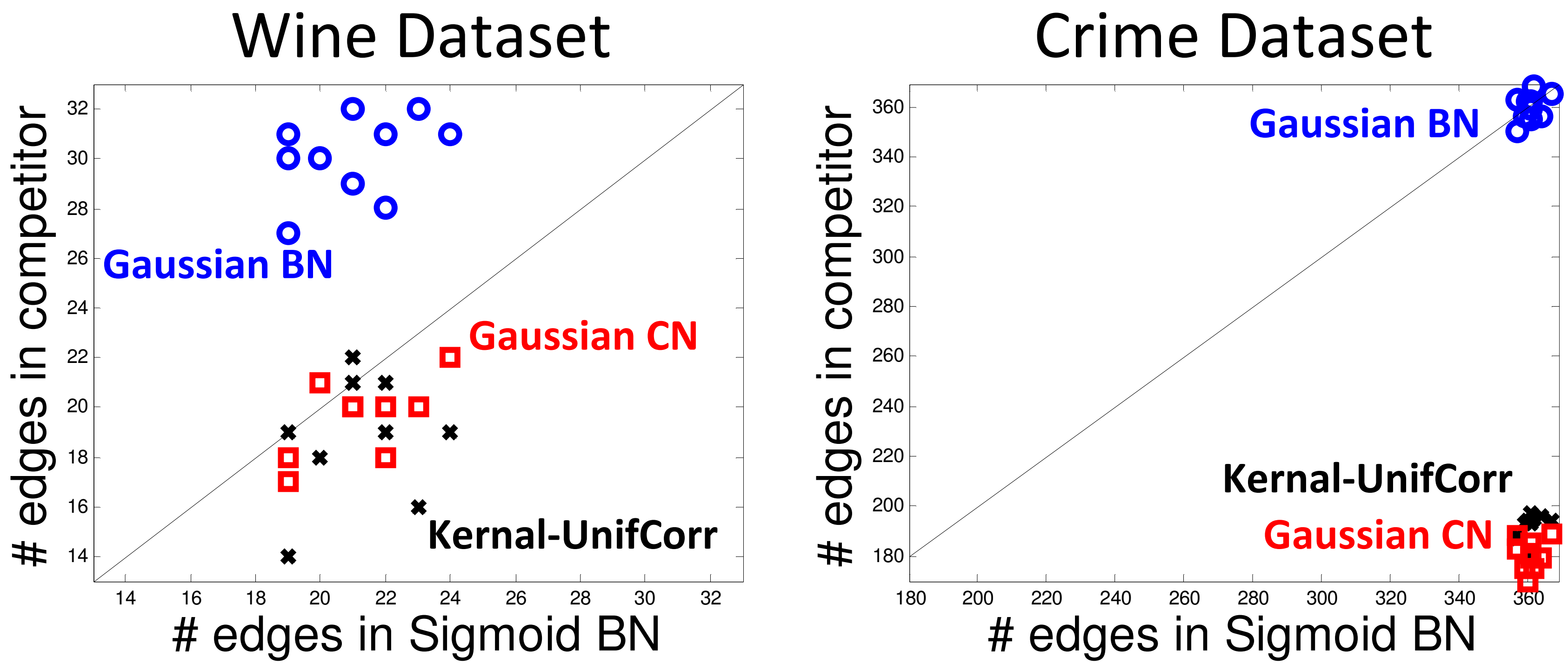
Crime (100 variables)



- ✓ Copula networks dominate BN models
- ✓ Learn structure in less than ½ hour!

Elidan, 2010

Complexity of Dependency Structure

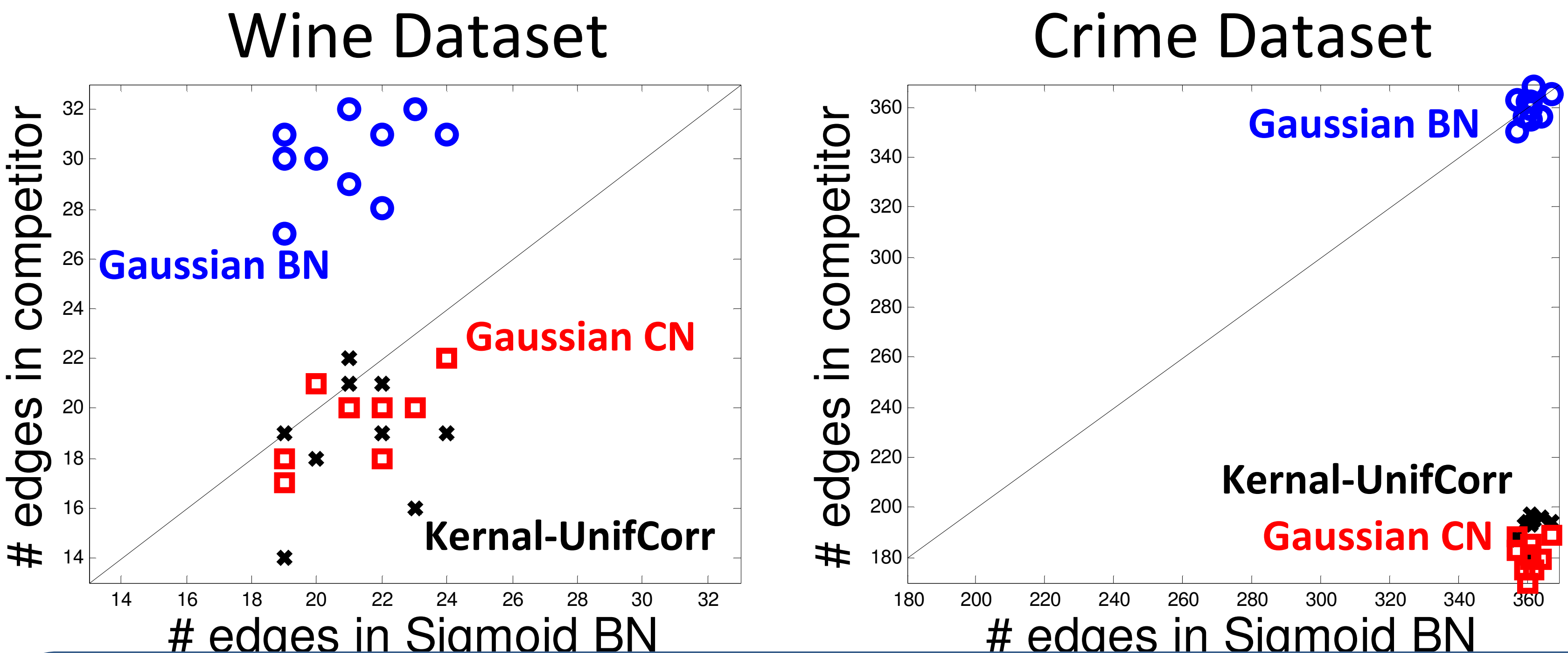


✓ Better generalization with sparser structures

✓ Simple (one parameter) copula resists over-fitting

Elidan, 2010

Complexity of Dependency Structure



Next steps: mean-field like inference (Elidan 2010) and lightning-speed structure learning (Elidan 2012)

Elidan, 2010

Real-life Processes

Motivation:

- Relationship between distance and velocity of rocket
- Relationship between volatilities of RVs, e.g. the returns on equity indices (hetero-scedastic sequence)

Challenges:

- Infinitely many interacting variables Z_t
- Non-Gaussian interaction
- Varied marginal distributions

Wilson and Ghahramani, 2010
See also related work by Jaimungal and Ng, 2009

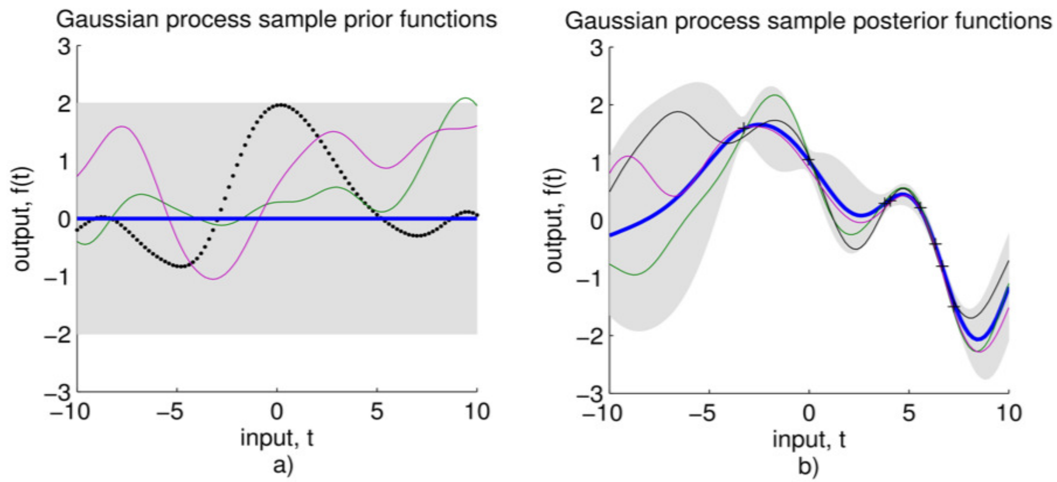
Gaussian Processes

A collection of random variables Z_t, any finite number of which have a joint Gaussian distribution

Used to define distribution over functions:

$$f(z) \sim \mathcal{GP}(m(z), k(z, z'))$$

1. any finite set $\{f(z_i)\}$ have a joint Gaussian distribution
2. $m(z_i)$ is the expectation of $f(Z_i)$
3. $\Sigma_{ij}=k(z_i,z_j)$ defines the functions properties



Rasmussen and Williams 2006 for (many) more details

Copula Processes

Let μ be a process measure with marginals G_t and joint H.

Z_t is a **copulas process** distributed with base measure μ if

$$P\left(\cap_{i=1}^{n}\left\{G_{t_i}^{-1}\left(F_{t_i}(Z_{t_i})\right) \leq a_i\right\}\right) = H_{t_1,\ldots,t_n}(a_1, \ldots, a_n)$$

Example: Gaussian Copula Process = μ is a standard GP

Another way to think about this:

There is a mapping Ψ that transform Z_t into a GP

$$\Psi(Z_t) \sim \mathcal{GP}(m(t), k(t, t'))$$

Wilson and Ghahramani, 2010

Gaussian Copula Process Volatility

Let $y_1,\ldots,y_n$ be a heteroscedastic sequence (varying σ_t)

Goal: model joint of $\sigma_1,\ldots,\sigma_n$ and predict unrealized σ_t

1. Observations: $y(t) \sim \mathcal{N}(0, \sigma^2(t))$ [this can be relaxed]
2. Volatility modeled as a Gaussian Copula Process

$$f(t) = \Psi^{-1}(\sigma(t)) \qquad \text{[warping function]}$$

$$f(t) \sim \mathcal{GP}(m(t) = 0, k(t, t'))$$

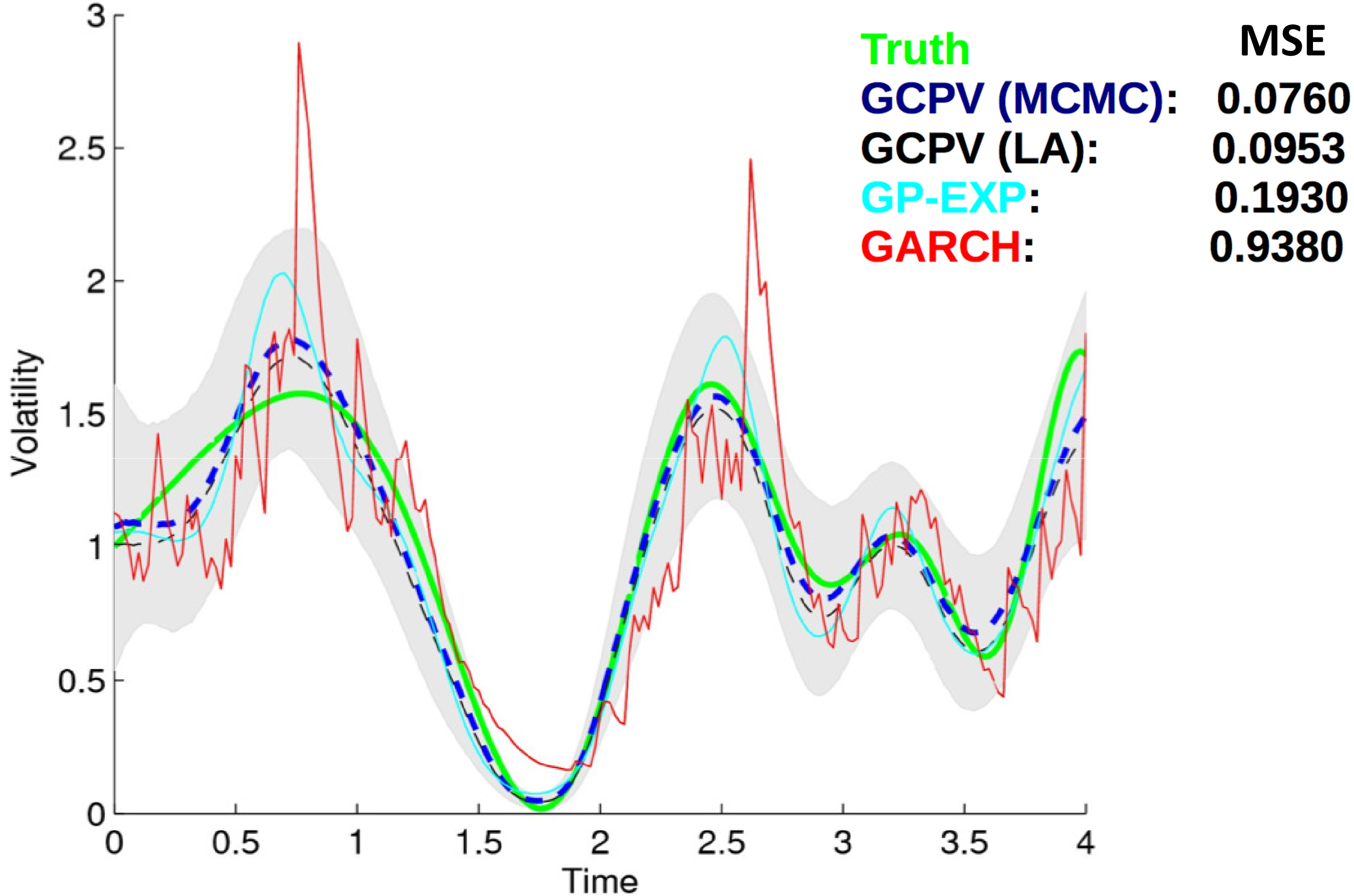
Challenges:

- Learn a flexible g (warping function)
- Need to do inference over many latent RVs

Interesting technical solutions in the paper! (no time ☹)

Wilson and Ghahramani, 2010

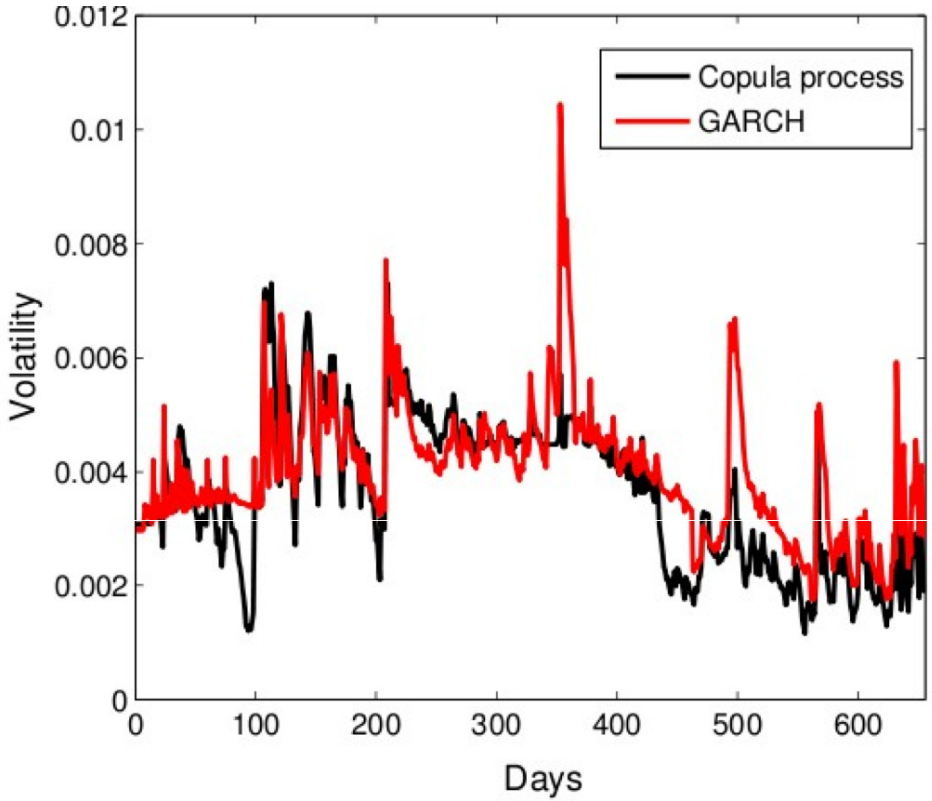
Simulation Results



Very promising results also for “JUMP” (spike like) sequence

Wilson and Ghahramani, 2010

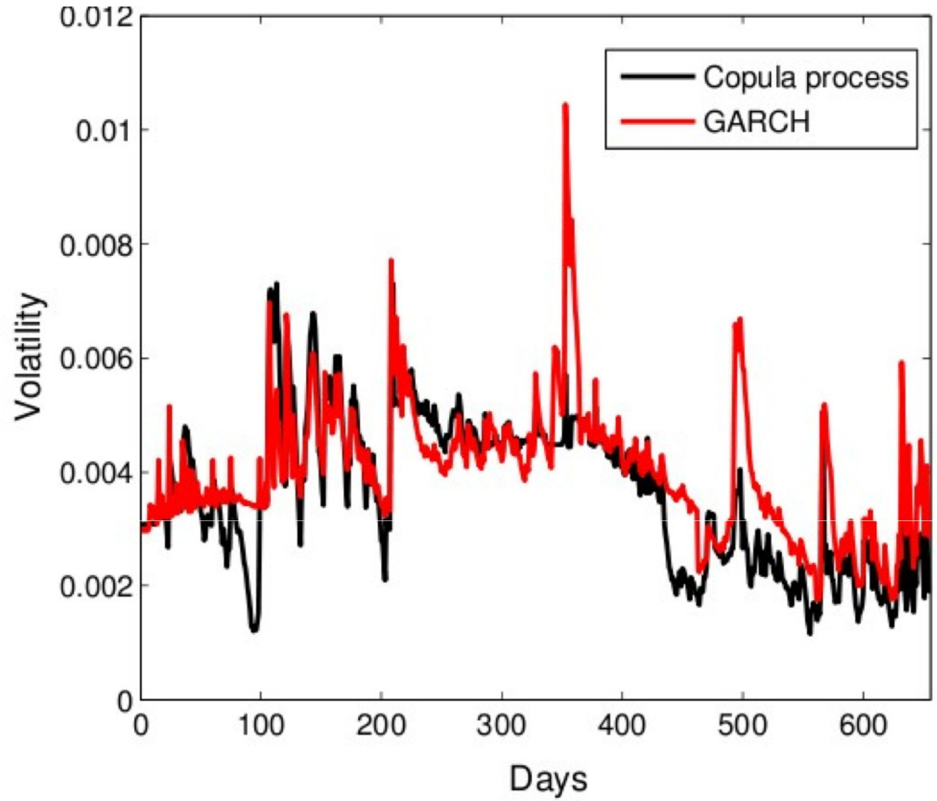
DM-GBP exchange rate returns



	Model	Historical	1 step	7 step	30 step
DMGBP	GCPV (LA)	2.43	3.00	3.08	3.17
	GCPV (MCMC)	2.39	3.00	3.08	3.17
$\times 10^{-9}$	GP-EXP	2.52	3.20	3.46	5.14
	GARCH	2.83	3.03	3.12	3.32

Wilson and Ghahramani, 2010

DM-GBP exchange rate returns



Next step: multivariate stochastic predictions

“Generalised Wishart Processes”, Wilson and Ghahramani 2011

Wilson and Ghahramani, 2010

Summary

Model	Base Copula	# RVs	Structure	Central merit
Vines	any bivariate	<10s	conditional dependence	Well understood general purpose framework
NPBBN	any bivariate	100s	BN+Vines	Mature application to large hybrid domains
Tree-averaged	any bivariate	10s	Markov	Bayesian averaging over structures
Non-paranormal	Gaussian	100-1000s	Markov	Large scale undirected estimation with guarantees
Copula Networks	any multivariate	100s	BN	General directed model that avoids conditional correlations
Copula Processes	any multivariate	∞ of few dimensions	-	Arbitrarily many variables

Further Information

- Vine Copula Handbook (Kurwicka and Joe, 2011)
- PhD thesis and papers on NPBBN (Hanea 2008,2009,2010)
- Tree-averaged distributions (Kirshner, 2008)
- The Nonparanormal (Liu, Wasserman and Lafferty, 2009)
- Copula BNs (Elidan, 2010), Inference-less Density Estimation (Elidan, 2010), Structure learning (Elidan, 2012)
- Copula Processes (Wilson and Ghahramani, 2010), Generalized Wishart Processes (W&G, 2011), Kernel-based Copula Processes (Jaimungal and Ng, 2009)