Heuristics for General Efficient Estimation in Copula Models and an Example

Peter J. Bickel

Department of Statistics
University of California at Berkeley

(with assistance of Jorge Bañuelos)

October 12, 2013

Joint work with Qunhua Li, James (Ben) Brown, Haiyan Huang

Table of Contents

The semiparametric copula model

$\mathcal{M} \equiv$ parametric "kernel" regular model
$\equiv \left\{ f(\cdot, \theta) \text{ densities on intervals} J \subset \mathbb{R}, \theta \in \Theta \text{ open in } \mathbb{R}^d \right\}$.

Semiparametric "copula" model:

$$\mathcal{C}_{SP}(\mathcal{M}) = \{ P_\theta \mathbf{T} \mid P_\theta \in \mathcal{M}, \mathbf{T} \in \mathcal{T} \}$$

$\mathbf{T} \equiv (T_1, \ldots, T_p)$, $T_j : \mathbf{J} \to J$, $\mathbf{J}$ an interval, $T_j' > 0$, $j = 1, \ldots, p$,
$\mathcal{T} \equiv$ all such transformations

Semiparametric copula (continued)

If $X = (X_1, \dots, X_p)$:

$$X \sim P_\theta \mathbf{T} \Leftrightarrow (T_1(X_1), \dots, T_p(X_p)) \sim f(\cdot, \theta).$$

WLOG in future $p = 2$.

Relation to copula generated by $\mathcal{M}$

$$\textbf{Copula} : \mathcal{C}(\mathcal{M}) = \left\{ P_\theta F^{-1}(\cdot, \theta) : \theta \in \Theta \right\}$$

where $F(\cdot, \theta), = (F_1(\cdot, \theta), F_2(\cdot, \theta))$ marginal cdf.
So,

$$X_j \sim U(0,1), \; j = 1, 2$$

Semiparametric copula (continued)

If $X^{(1)}, \ldots, X^{(n)}$ iid $\mathcal{C}_{SP}(\mathcal{M})$, $\hat{R}^{(i)} = (R_{1i}, R_{i2})$, and $R_{ij} = \sum_{k=1}^{n} 1(X_j^{(k)} \leq X_j^{(i)})$ then for $(p = 2)$, $\hat{R}^{(1)}, \ldots, \hat{R}^{(n)}$ are asymptotically sufficient, , LAN. See also Hoff (2007), and Bickel, Ritov (1997).

Basic Questions

Given an identifiable parametrization, $(\theta, \mathbf{T})$:

$$P_{\theta_1, \mathbf{T}_1} = P_{\theta_2, \mathbf{T}_2} \Leftrightarrow \theta_1 = \theta_2, \mathbf{T}_1 = \mathbf{T}_2$$

1. Construct a fitting algorithm which converges for fixed n:

$$(\hat{T}_m, \hat{\theta}_m) \to (\hat{T}, \hat{\theta})$$

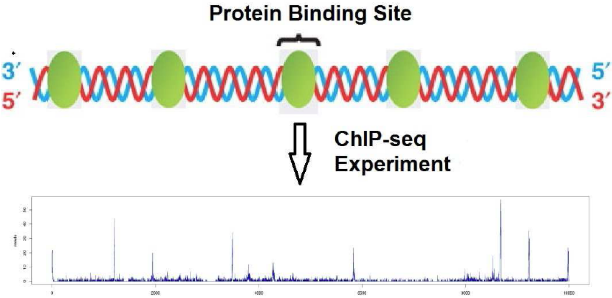
2. Have $(\hat{\mathbf{T}}, \hat{\theta})$ which are semiparametrically efficient:

$$(\sqrt{n}(\hat{\theta} - \theta_0), \sqrt{n}(\hat{\mathbf{T}} - \mathbf{T}_0))$$

   are asymptotically Gaussian, and achieving the information bound.
   Treat problem for $\mathcal{M}$ and $\mathcal{C}$ as equivalent, assuming $\theta \to F(\cdot, \theta)$ smooth.

Measuring reproducibility of high-throughput experiments
Qunhua Li, James B. Brown, Haiyan Huang, Peter J. Bickel
Annals of Applied Statistics, Volume 5, Number 3
(2011), 1752 - 1779.

A Motivating Example: ChIP-seq experiment



Signal Identification

Identified Peaks			significance score
chromosome	start	end	
chr1	713885	716999	257.05
chr22	21614704	21616162	238.75
chr9	132731100	132732425	234.01
chr10	135372185	135373701	217.76
chr5	43141138	43142110	215.97
chr16	33859768	33867019	208.77
...	...	...	...
...	...	...	...

- Significance value represents relative strength of the signal. Commonly used: fold of enrichment, p-value, q-value
- Significance scores usually are not on well-calibrated probabilistic measures
  - The null distribution is difficult to approximate
  - Scale may vary across datasets

Arbitrary judgement often is involved in the selection of threshold

ENCODE's request

Goal: *Uniformly* process data from *multiple sources*.

- Compare performance of different algorithms and select the best one to process data
- Select peaks using a uniform criterion across all datasets

However,

- No ground truth is available

Can replicates help?

Genuine signals should be reproducible across replicates

Can we use replicates to

- select reproducible signals?
- assess the reproducibility of algorithms?

Consistency between calls on two replicates

Suppose X and Y are the significance values on two replicates.

- Assume X and Y reasonably reflect relative strength of signals
- Their distributions are *unknown* and may be *different*.

	Rank(X)	X	Y	Rank(Y)	
	100	10	31	100	
	99	9	30	99	good
Signal	$\boxed{98}$	8	27.9	$\boxed{97}$	agreement
	$\boxed{97}$	7.5	28.1	$\boxed{98}$	
	96	7.4	27.5	96	
		...	...		
		...	...		
	$\boxed{3}$	0.6	10.7	$\boxed{15}$	
Noise	$\boxed{2}$	0.5	10.8	$\boxed{20}$	bad
	$\boxed{1}$	0.4	11.1	$\boxed{30}$	agreement

- Correspondence is expected to decay when getting to noise
- The divergence point provides a guidance on how many calls cannot be trusted

Encode Data

- ChIP-seq experiments on transcription factor CTCF (Broad Institute)
- 2 biological replicates
- 9 algorithms
  - Enrichment: Erange, Fseq, QuEst, SPP, Cisgenome
  - p-value: HotSpot, MACS, SISSRS
  - q-value: Peakseq
- Peaks are normalized to a unified width

A copula mixture model

$p = 2$

- X_{ij}: Intensity of Peak i on replicate j, $j = 1, 2$.
- Status of peaks

$$S_i = \begin{cases} 1 & \text{if reproducible} \\ 0 & \text{if irreproducible} \end{cases}$$

- Assume the dependence in each component is induced from a Gaussian distribution (z_0 and z_1) with different association parameters ($\rho_0 = 0, \rho_1 > 0$).
- Assume z_1 is stochastically larger than z_0, i.e. $\mu_1 > \mu_0$, $\mu_0 = 0$.

Statistical Model

Let $(X_{i1}, X_{i2})_{i=1,\ldots,n}$ = Intensity of peak after 2 replicates
Assume/Pretend:

1. These behave like a sample from a population.
2. On possibly different scales

$$T_1, T_2 : \mathbb{R} \to \mathbb{R}, \uparrow \textit{differentiable},$$

   a peak pair is distributed as a mixture of 2 bivariate Gaussian distribution

Copula Model

$$(T_1(X_{11}), T_2(X_{12})) \sim (1-\epsilon)N(\mu, \mu, \sigma^2, \sigma^2, \rho) + \epsilon N(0, 0, 1, 1, 0)$$

with $\mu > 0$, $\theta = (\mu, \mu, \sigma^2, \sigma^2, \rho)$ unknown.

$$N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho) \equiv \text{ bivariate Gaussian distribution}$$

Statistical Model

$$\text{I} = \text{“good peaks”}, \quad \text{II} = \text{“noisy peaks”}$$

Scales (T_1, T_2) are unknown so model is

$$\begin{aligned} P(X_{11} < x, X_{12} < y) &= (1-\epsilon)\Phi\left(T_1^{-1}(x), T_2^{-1}(y); \theta\right) \\ &+ \epsilon\Phi\left(T_1^{-1}(x)\right)\Phi\left(T_2^{-1}(y)\right) \end{aligned}$$

where $\theta = (\mu, \mu, \sigma^2, \sigma^2, \rho)$

Irreproducible discovery rate

In analogy to multiple testing,

- Two groups: *Reproducible* vs. *Irreproducible*
- Local irreproducible discovery rate

$$\mathbf{idr}(x_1, x_2) = \frac{\pi_0 f_0(x_1, x_2)}{f(x_1, x_2)}$$

- Irreproducible discovery rate

$$\mathbf{IDR}(\gamma) \equiv P(\text{irreproducible} \mid I_\gamma) = \frac{\pi_0 \int_{I_\gamma} dF_0(x_1, x_2)}{\int_{I_\gamma} dF(x_1, x_2)}$$

where $I_\gamma = \{(x_1, x_2) : \mathbf{idr}(x_1, x_2) < \gamma\}$.

Irreproducible discovery rate

- For a desired control level α, define $\gamma_0 = \arg\max_\gamma\{\mathbf{IDR}(\gamma) \le \alpha\}$. Selecting all pairs $\in I_{\gamma_0}$ gives an expected rate of irreproducible discoveries no greater than α.
- Selection is based on likelihood ratio, different from thresholding based on significance scores

Abridged References

1. Recent complete exhaustive treatment of Gaussian copulas using "1-step" estimator: Segers, Van der Akker, Werber (2013) following Klassen, Wellner (1997), Hoff, Niu, Wellner (2012).
2. Pseudo-likelihood: Segers, Van der Akker, Werber (2013) following Klassen, Wellner (1997), Hoff, Niu, Wellner (2012), Genest, Rivest (1993) + many others
3. Sieves - General & Efficient but many tuning parameters: Chen, Fan, Tsyrennikov (2006).

Our Method of Fitting

$\mathcal{M}$: $f(x, y, \theta)$
$\mathcal{C}$:

$$\frac{f(F_1^{-1}(u,\theta), F_2^{-1}(v,\theta), \theta)}{f_1(F_1^{-1}(u,\theta)) f_2(F_2^{-1}(v,\theta))}$$

Given: $(\hat{F}_1(x_i), \hat{F}_2(y_i))$, $i = 1, \ldots, n$.

- Form $(x_i(\theta), y_i(\theta))$, where

$$x_i(\theta) = F_1^{-1}(\hat{F}_1(x_i), \theta), y_i(\theta) = F_2^{-1}(\hat{F}_2(y_i), \theta)$$

- Maximize

$$\sum_{i=1}^{n} \log f(x_i(\theta), y_i(\theta), \theta)$$

Advantage: Can use EM for Mixture, NO Tuning Parameters

Method of Fitting

Our method (when scaled):

- Works very well in practice: Ibrahim, Daniel M., et al. (2013), Gao, Xia, et al. (2013), + Others
- Heuristics for $\sqrt{n}$-consistency
- Not efficient
- No algorithm convergence proven

Proposed Algorithm Revision

A revision sharing same features but heuristically *efficient*. Argument for $p = 2$ but simply generalizable. For simplicity write $X_2 = Y$.

- An "NPMLE" for T_1, T_2, θ_0 fixed.

$$\mathcal{L} \equiv \int \left(\ell(T_1(x), T_2(y), \theta_0) + \log T_1'(x) + \log T_2'(y)\right) d\hat{F}(x, y)$$

- $\ell \equiv$ loglikelihood for $n = 1$, parametric model.
- $\hat{F} \equiv$ Empirical distribution of (X_i, Y_i), $i = 1, \ldots, n$ i.i.d. F_{θ_0}.

WLOG $T_1^0, T_2^0 =$ Identity

Variational optimization (and formal trick)

1. Reparametrize $a_1(x) \equiv \log T_1'(x)$, $a_2(y) \equiv \log T_2'(y)$.
2. Replace $d\hat{F}(x,y)$ by $f(x,y)\ dx\ dy$, f arbitrary

Then

$$T_1(x) = \int_0^x \exp\{a_1(u)\}\ du,\ \ T_2(y) = \int_0^y \exp\{a_2(v)\}\ dv$$

Take $J = [0,1]$.

Variational optimization (and formal trick)

Let

$$a_{1\epsilon}(u) = a_{10}(u) + \epsilon\triangle_1(u),\ \ a_{2\epsilon}(v) = a_{20}(v) + \epsilon\triangle_2(v)$$

where a_{10}, a_{20} maximizers for θ_0, f fixed.
For $a_1 = a_{1\epsilon}, a_2 = a_{2\epsilon}$,

$$\mathcal{L}_\epsilon(T_{1\epsilon}, T_{2\epsilon}) \equiv \mathcal{L}$$

Variational argument (continued)

$$\frac{\partial \mathcal{L}}{\partial \epsilon}(a_{10}, a_{20}) \mid_0 = \int_0^1 \int_0^1 \left\{ \int_u^x \left[e^{a_{10}(u)} \triangle_1(u) \; du \right] \ell_1(T_{10}(x), T_{20}(y), \theta_0) + \int_v^y \left[e^{a_{20}(v)} \triangle_2(v) \; dv \right] \ell_2(T_{10}(x), T_{20}(y), \theta_0) + \triangle_1(x) + \triangle_2(y) \right\} f(x, y) \; dx \; dy$$

for all $\triangle_1, \triangle_2$.
$\ell_1, \ell_2 \equiv$ partials with respect to first and second coordinates
$(a_{10}, a_{20}), (T_{10}, T_{20})$ maximizers

Variational argument (continued)

ZEROS OF DERIVATIVE OF $\mathcal{L}_\epsilon$

$$T'_{10}(u)\int_0^1\int_u^1 \ell_1(T_{10}(x), T_{20}(y), \theta_0) f(x,y)\, dx\, dy + f_{X_1}(u) = 0$$

and analogously for $T_{20}(v)$.

Variational argument (continued)

"NPMLE":
Integrate both sides *and plug in* $d\hat{F}$ for $f(x,y)\,dx\,dy$.

$$\hat{T}_1(x,\hat{F}) = -\int_0^x d\hat{F}_{X_1}(u)\Lambda_1^{-1}(\hat{T}_1,\hat{T}_2,\hat{F},\theta_0)(u) \tag{1}$$

where

$$\Lambda_1(\hat{T}_1,\hat{T}_2,\hat{F},\theta_0)(u) = \int_0^1\int_u^1 \ell_1(\hat{T}_1(s),\hat{T}_2(t),\theta_0)d\hat{F}(s,t)$$

Compute analogously for $\hat{T}_2(y,\hat{F})$ and label the resulting equation by (7).

Proposed Algorithm

1. Initialize with $\hat{\theta}_0$.
2. Let $\hat{T}_1(\cdot, \hat{\theta}_1)$, $\hat{T}_2(\cdot, \hat{\theta}_1)$ solve (6), (7)
3. At stage m let $X_i(\hat{\theta}_{m-1}) = \hat{T}_1(X_i, \hat{\theta}_{m-1}),\ \ Y_i(\hat{\theta}_{m-1}) = \hat{T}_2(Y_i, \hat{\theta}_{m-1})$
4. Let $\hat{\theta}_m$ maximize
$$\sum_{i=1}^{n} \ell\left[\hat{T}_1(X_i(\hat{\theta}_{m-1})), \hat{T}_2(Y_i(\hat{\theta}_{m-1})), \theta\right]$$
5. Determine $\hat{T}_1(X_i(\hat{\theta}_m))$, $\hat{T}_2(Y_i(\hat{\theta}_m))$, by solving (6), (7)
6. Repeat until convergence

That is, follow Li et al. (2011) but $\hat{T}_1(X_i(\hat{\theta}_m))$ replaces $F_1^{-1}(\hat{F}_1(X_i), \hat{\theta}_{m-1})$ and similarly for $\hat{T}_2$.
Assume algorithm converges to $(\hat{\theta}, \hat{\mathbf{T}})$ where $\hat{\mathbf{T}}$ satisfies (6), (7) and $\hat{\theta}$ the likelihood equations.

Issues with Proposed Algorithm

1. No obvious way of solving (6), (7)
2. Solutions may not be monotone $\uparrow$. Certainly not $\uparrow$ strictly.

Heuristic Asymptotic Analysis of (6), (7)

The true $(\theta_0, \mathbf{T}_0)$ satisfies (6), (7).

$$\mathbf{T}_0(x, y, \theta_0) = (x, y) = \left(-\int_0^x \frac{dF_x(u, \theta_0)}{\Lambda_1(x, y, \theta_0)}, -\int_0^y \frac{dF_y(v, \theta_0)}{\Lambda_2(x, y, \theta_0)} \right)$$

Since

$$\begin{aligned} \Lambda_1(x, y, \theta_0)(u) &= \int_0^1 \int_u^1 \left[\frac{\partial f}{\partial x}(s, t, \theta_0) f^{-1}(s, t, \theta_0) \right] \dots \\ &\quad \dots f(s, t, \theta_0)\; ds\; dt \\ &= -f_X(u, \theta_0). \end{aligned}$$

Similarly, $\Lambda_2(x, y, \theta_0) = -f_Y(v, \theta_0)$.

Computation of Influence Function

Simplify: Assume that underlying distribution satisfies the copula assumption. If θ is true that means just redefining

$$X_i^{NEW}(\theta) = F_X(X_i, \theta), \ \ Y_i^{NEW}(\theta) = F_Y(Y_i, \theta)$$

Call the new transformations $T_{1\theta} = F_X T_1(X)$, $T_{2\theta} = F_Y T_2(Y)$.
So the truth is $(\theta_0, \mathbf{T}_{\theta_0})$ and the estimates are $\hat{\theta}, \hat{\mathbf{T}}_{\hat{\theta}}$.
We compute formally the influence function of $(\hat{\theta}, \mathbf{T})$ by expanding around $(\theta_0, \mathbf{T}_{\theta_0})$.

Computation of Influence Function (cont.)

Since

$$\Lambda_1(\mathbf{T}_{\theta_0})(u) = -f_x(u,\theta_0) = 1 = \Lambda_2(\mathbf{T}_{\theta_0})(v)$$

Then

$$\begin{aligned}\left[\Lambda_1(\hat{\mathbf{T}}_{\hat{\theta}}, \hat{F}) - \Lambda_1(\mathbf{T}_{\theta_0}, F_{\theta_0})\right](u) &= \int_0^1\int_0^1 \ell_x(x,y,\theta_0)dF_{\theta_0}(x,y) \\ &+ \int_0^1\int_0^1 \Big[\ell_{xx}(x,y,\theta_0)(\hat{T}_1 - T_{10})(x) \\ &+ \ell_{xy}(x,y,\theta_0)(\hat{T}_2 - T_{20})(y)\Big] \\ &\quad dF_{\theta_0}(x,y) \\ &+ \text{lower order terms}\end{aligned}$$

Where $\ell_x, \ell_{xx}, \ell_{xy}$ are partial second derivatives.
Similar expansion holds for $\Lambda_2(\hat{\mathbf{T}}_{\hat{\theta}}, \hat{F}) - \Lambda_2(\mathbf{T}_{\theta_0}, F_{\theta_0})$

Computation of Influence Function (cont.)

Thus, if we view $\mathbf{T}'s$ as members of BV(J)$\times$ BV(J),

$$(I - M)(\hat{\mathbf{T}}_{\hat{\theta}} - \mathbf{T}_{\theta_0})(x,y) = (A_1(x), A_2(y)) + \text{ lower order terms}$$

$M \equiv$ linear operator and

$$\begin{aligned} A_1(x) &= \int_0^1 1(x \geq u)d\ (\hat{F}_x(v,\theta_0) - v) \\ &+ \int_0^1 \int_0^1 1(x \geq u)\ell_x(u,v,\theta_0)\ d(\hat{F}_X(u,v) - F_{\theta_0}(u,v)) \end{aligned}$$

$$\begin{aligned} A_2(y) &= \int_0^1 1(y \geq v)d\ (\hat{F}_y(v,\theta_0) - v) \\ &+ \int_0^1 \int_0^1 1(y \geq v)\ell_y(u,v,\theta_0)\ d(\hat{F}(u,v) - F_{\theta_0}(u,v)) \end{aligned}$$

Computation of Influence Function (cont.)

Since

$$1(x \geq u) + 1(x \geq u)\ell_x(u, v, \theta_0), 1(y \geq v) + 1(y \geq v)\ell_y(u, v, \theta_0) \in \dot{P}_{\theta_0},$$

$\dot{P}_{\theta_0} \equiv$ tangent space with θ_0 fixed,

$$I - M \text{ nonsingular } \Rightarrow \hat{\mathbf{T}}_{\hat{\theta}} \text{ efficient}$$

Open questions

1. Consistency of $(\hat{\theta}, \hat{\mathbf{T}})$
2. Invertibility of $I - M$
3. Validity of expansions
4. No guarantee that $\hat{\mathbf{T}}$ is increasing.
5. It may be necessary for simplification to replace $d\hat{F}$ by $\hat{f}(x, y)\, dx\, dy$ i.e. smooth.

Possible Answers

1-3. If $\hat{\theta}_0$ is close enough to θ_0 and $\hat{\mathbf{T}}_{\hat{\theta}}$ to $\mathbf{T}_{\theta_0}$ Cramér's approach and Banach fixed point theorem should work.

4. Approximate each iteration by monotone function; Replace $\hat{T}_1(X_{(i)}) - \hat{T}_1(X_{(i-1)})$ by its positive part.

- **Conjecture**: If we can construct an estimate of θ, $\hat{\theta}$, that is consistent and converges uniformly to θ_0 then this construction using $\hat{\theta}$ and $(F_X^{-1}(\cdot, \hat{\theta})\hat{F}_X, F_Y^{-1}(\cdot, \hat{\theta})\hat{F}_Y)$ as a starting point should give efficiency. Constructions using pseudolikelihood should do it.

How broadly applicable should it be?

- Extension to $p > 2$ should work in principle although computation may be burdensome

Relation to Chen, Fan, Tsyrennikov (2006)

Approach similar for smoothed proposal. BUT:

1. We distinguished between θ and $\hat{\mathbf{T}}$ using fact that fitting θ for $\hat{\mathbf{T}}$ fixed may have familiar algorithm.
2. Our argument suggests nothing special about splines over other smoothing.
3. If our original approach works no tuning parameters needed