Adaptive Nonparametric Likelihood Weights

Jean-François Plante
University of Toronto

From of a Doctoral thesis completed under the supervision of
James V. Zidek
at the University of British Columbia

Workshop on Composite Likelihood Methods, University of Warwick
17 April 2008

Plan

- The weighted likelihood
- The Entropy Maximization Principle
- The MAMSE weights
  - Definition
  - Properties
  - Simulation results
- Other applications of the heuristics

The Weighted Likelihood

Available data: $X_{ij} \overset{\perp\!\!\!\perp}{\sim} F_i,$ $\quad i = 1, \ldots, m$ (Population), $\quad j = 1, \ldots, n_i$ (Individual)

- Samples come from m populations
- Inference is about Population 1
- The family of distributions $f(x|\theta)$ is used to model Population 1

$$L_{\boldsymbol{\lambda}}(\theta) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} f(X_{ij}|\theta)^{\lambda_i/n_i}$$

The Maximum Weighted Likelihood Estimate (MWLE) is a value of θ maximizing $L_{\boldsymbol{\lambda}}(\theta)$.

The *weighted log-likelihood* may be more intuitive:

$$\ell_{\boldsymbol{\lambda}}(\theta) = \sum_{i=1}^{m} \frac{\lambda_i}{n_i} \sum_{j=1}^{n_i} \log f(X_{ij}|\theta)$$

The weights discount data based on their relevance (or lack thereof).

How to choose them?

- Scientific information.
- Ad-hoc method: Hu & Zidek (2002).
- Cross-validation: Wang & Zidek (2005).

None of these solutions is fully satisfactory.

Note: The paradigm adopted is the same as Wang (2001) and Wang & Zidek (2005).

Maximum Entropy and Maximum Likelihood

Maximum Entropy Principle

In the family $f(x|\theta)$, choose the distribution closest to f_1 (the true distribution) by maximizing the Entropy:

$$\begin{aligned} B(f_1, f) &= -\int \frac{f_1(x)}{f(x|\theta)} \log\left\{\frac{f_1(x)}{f(x|\theta)}\right\} f(x|\theta)\,\mathrm{d}x \\ &= \int \log\{f(x|\theta)\} f_1(x)\,\mathrm{d}x - \int \log\{f_1(x)\} f_1(x)\,\mathrm{d}x \end{aligned}$$

We can ignore the second term because it does not depend on θ.

But f_1 is unknown! What to do?

Suggestion #1: Use the empirical CDF

$$\hat{F}_1(x) = \frac{1}{n_1} \sum_{j=1}^{n_1} \mathbb{1}(X_{1j} \le x)$$

as a "good guess" for the true distribution.
$\hat{F}_1(x)$ allocates a weight of $1/n_1$ to each data point.

$$\text{Entropy} \sim \int \log f(x|\theta)\,\mathrm{d}\hat{F}_1(x) = \frac{1}{n_1} \sum_{j=1}^{n_1} \log f(X_{1j}|\theta),$$

the log-likelihood!!!

Suggestion #2: Use a mixture of m empirical CDF's

$$\hat{F}_{\boldsymbol{\lambda}}(x) = \sum_{i=1}^{m} \lambda_i \hat{F}_i(x) \qquad \text{with } \lambda_i \geq 0 \text{ and } \sum_{i=1}^{m} \lambda_i = 1.$$

Each data point has weight λ_i / n_i.

Then,

$$\text{Entropy} \sim \int \log f(x|\theta) \, \mathrm{d}\hat{F}_{\boldsymbol{\lambda}}(x) = \sum_{i=1}^{m} \frac{\lambda_i}{n_i} \sum_{j=1}^{n_i} \log f(X_{ij}|\theta),$$

the weighted log-likelihood!!!

Intuitively, weighted likelihood $\sim$ using $\hat{F}_{\boldsymbol{\lambda}}$ to estimate F_1.

The MAMSE Weights

Based on the previous heuristic development, we want:

- $\hat{F}_{\boldsymbol{\lambda}}$ close to F_1
- $\hat{F}_{\boldsymbol{\lambda}}$ less variable than $\hat{F}_1$

We combine these requirements into an objective function:

$$P_{\boldsymbol{\lambda}} = \int \left[\left\{ \hat{F}_1(x) - \hat{F}_{\boldsymbol{\lambda}}(x) \right\}^2 + \widehat{\mathrm{var}} \left\{ \hat{F}_{\boldsymbol{\lambda}}(x) \right\} \right] \mathrm{d}\hat{F}_1(x)$$

where $\widehat{\mathrm{var}}\{\hat{F}_{\boldsymbol{\lambda}}(x)\} = \sum_{i=1}^{m} \frac{\lambda_i^2}{n_i} \hat{F}_i(x)\{1 - \hat{F}_i(x)\}$.

We choose the weights that minimize $P_{\boldsymbol{\lambda}}$ and call them MAMSE (Minimum Averaged Mean Squared Error) weights.

$$P_{\boldsymbol\lambda} = \int \left[\left\{ \hat{F}_1(x) - \hat{F}_{\boldsymbol\lambda}(x) \right\}^2 + \sum_{i=1}^{m} \frac{\lambda_i^2}{n_i} \hat{F}_i(x)\{1 - \hat{F}_i(x)\} \right] \mathrm{d}\hat{F}_1(x)$$

Note that:

- $P_{\boldsymbol\lambda}$ is quadratic in $\boldsymbol\lambda \Rightarrow$ easy to optimize.
- $P_{\boldsymbol\lambda}$ does not depend on the model $f(x|\theta)$.
- The MWLE is invariant to a reparametrization $f(x|\theta)$.

By their definition, the MAMSE weights can be used as:

- Likelihood weights.
- Mixing probabilities for the empirical functions $\hat{F}_i(x)$.

Asymptotics

Consider a sequence of samples such that $n_1 \to \infty$.
We assume the distributions (F_i) are continuous.
Then,

- "Glivenko-Cantelli":

$$\sup_x \left| \hat{F}_{\boldsymbol\lambda}(x) - F_1(x) \right| \to 0 \quad a.s.$$

- Strong Law of Large Numbers: for a suitable function g,

$$\sum_{i=1}^{m} \frac{\lambda_i}{n_i} \sum_{j=1}^{n_i} g(X_{ij}) \to E\{g(X_1)\} \quad a.s.$$

- Suppose that $F_1(x) \equiv F(x|\theta_0)$, then
  the MWLE is a strongly consistent estimator of θ_0.

A Word About the Proof of Consistency

Adapted from Wald (1949).

- For any θ outside an open set containing θ_0,
  the likelihood is bounded.
- Use some properties of the Relative Entropy.
- Critical point: must have a Strong Law of Large Numbers

$$\int \log f(x|\theta)\,\mathrm{d}\hat{F}_{\boldsymbol{\lambda}}(x) \to \int \log f(x|\theta)\,\mathrm{d}F_1(x)$$

  LHS = log-likelihood,
  RHS = expectation under the true model.

Simulations

1. Normal Distribution

Samples of size n from each of

$$\begin{aligned} \text{Population 1} &: \mathcal{N}(0,1) \\ \text{Population 2} &: \mathcal{N}(\Delta,1) \end{aligned}$$

Number of replicates: 10000.

Note: Results hold for $\mathcal{N}(\mu,\sigma^2)$ and $\mathcal{N}(\mu+\Delta\sigma,\sigma^2)$ as well.

	Average Value of $100\lambda_1$					
	$n=10$	20	50	100	1000	10000
$\Delta=0$	71	71	72	72	72	72
0.01	72	72	72	72	72	74
0.10	72	73	73	74	86	98
0.25	74	76	79	83	97	100
0.50	79	82	88	93	99	100
1.00	87	92	96	98	100	100
2.00	94	97	99	99	100	100

TABLE 1. Average MAMSE weights for Simulation 1, Pop. 1: $\mathcal{N}(0,1)$, Pop. 2: $\mathcal{N}(\Delta,1)$.

	100 MSE(MLE)/MSE(MWLE)					
	$n=10$	20	50	100	1000	10000
$\Delta=0$	145	144	143	144	144	143
0.01	146	144	143	144	141	127
0.10	143	140	135	128	89	94
0.25	134	125	110	96	91	99
0.50	117	104	88	88	97	100
1.00	94	88	90	94	99	100
2.00	87	92	96	98	100	100

TABLE 2. Relative efficiency for Simulation 1, Pop. 1: $\mathcal{N}(0,1)$, Pop. 2: $\mathcal{N}(\Delta,1)$.

2. Complementary Populations

Samples of size n are drawn from

$$\begin{aligned} \text{Population 1} &: \mathcal{N}(0,1) \\ \text{Population 2} &: |\mathcal{N}(0,1)| \\ \text{Population 3} &: -|\mathcal{N}(0,1)|. \end{aligned}$$

Each scenario is repeated 10000 times.

n	Efficiency	$100\bar{\lambda}_1$	$100\bar{\lambda}_2$	$100\bar{\lambda}_3$
10	121	46	23	30
20	118	45	25	29
50	117	45	27	29
100	116	44	27	28
1000	115	44	28	28
10000	116	44	28	28

Table 3. Average weights and efficiency for Simulation 2. Pop. 1: $\mathcal{N}(0,1)$, Pop. 2: $|\mathcal{N}(0,1)|$, Pop. 3: $-|\mathcal{N}(0,1)|$.

Efficiency=100 MSE(MLE)/MSE(MWLE)

Using the Similar Heuristics in Other Contexts

Suppose $\mathbf{x}$ as q dimensions. Working assumption: all elements of $\mathbf{x}$ are independent except x_i and x_j. Then, $\int \log f(\mathbf{x}|\theta)\,\mathrm{d}\hat{F}(\mathbf{x})$ equals

$$\int \log f(x_i, x_j|\theta)\,\mathrm{d}\hat{F}(\mathbf{x}) + \sum_{k \notin \{i,j\}} \int \log f(x_k|\theta)\,\mathrm{d}\hat{F}(\mathbf{x})$$

Consider this assumption for all possible pairs of variables.
We compromise by maximizing their sum $\Rightarrow$
a composite likelihood in the sense of Cox & Reid (2004).

For a single observation:

$$\sum_{i<j} \log f(x_i, x_j|\theta) + \binom{q-1}{2} \sum_{i=1}^{q} \log f(x_i|\theta).$$

Could this be useful ?

Conclusion

A heuristic justification of the weighted likelihood leads to the definition of the MAMSE weights.

The nonparametric MAMSE weights yield consistent estimates that allow to borrow strength from other populations without making parametric assumptions on them.

The MAMSE weights are useful in other contexts too
(survival analysis, nonparametric coefficients of correlation, copulas).

The heuristic used for the weighted likelihood may be useful in developing further other composite likelihoods...

Thank You!

References

D. R. Cox & N. Reid (2004). A note on pseudolikelihood constructed from marginal densities, *Biometrika*, **91**, 729–737.

F. Hu and J. Zidek (2002). The weighted likelihood, *The Canadian Journal of Statistics*, **30**, 347–371.

A. Wald (1949). Note on the consistency of the maximum likelihood estimate, *The Annals of Mathematical Statistics*, **20**, 595–601.

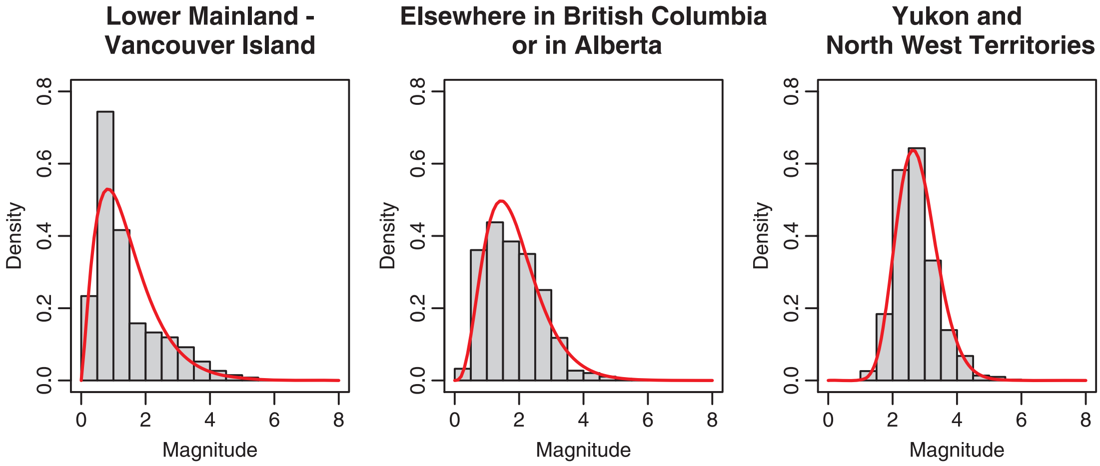
X. Wang (2001). *Maximum weighted likelihood estimation*, unpublished doctoral dissertation, Department of statistics, University of British Colombia, 151 pp.

X. Wang and J. V. Zidek (2005). Selecting likelihood weights by cross-validation, *The Annals of Statistics*, **33**, 463–501.

Additional Simulation

3. Earthquakes

Magnitude of earthquakes in Western Canada in a five-year period. Data from the public website of *Natural Resources Canada*.



Number of measured earthquakes: 4743, 4866 and 1621 respectively. The red line corresponds to the fitted Gamma distribution.

Suppose that the magnitude of the 50 last earthquakes from each area are available. Should we use the MLE or the MWLE?

We generated 10000 samples of 50 earthquakes from each area based on the fitted model. We calculated the MLE and the MWLE for the Lower Mainland – Vancouver Island area for each sample.

Average weights: 0.959, 0.041, 0.000 respectively.

Estimation of the parameters:

$$100\ \mathrm{MSE(MLE)/MSE(MWLE)}=107$$

Estimation of P(Magnitude $>$ 3):

$$100\ \mathrm{MSE(MLE)/MSE(MWLE)}=112$$

Copulas

Suppose m samples of p-dimensional data are available from continuous distributions:

$$\mathbf{X}_{ij} = (X_{ij1}, \ldots, X_{ijp}) \overset{\perp\!\!\!\perp}{\sim} F_i, \qquad \begin{array}{ll} i = 1, \ldots, m & \text{(Population)} \\ j = 1, \ldots, n_i & \text{(Individual)} \end{array}$$

with $\qquad F_i(\mathbf{x}) = C_i\{G_{i1}(x_1), \ldots, G_{ip}(x_p)\}$

where C_i is the unique copula associated with F_i and $G_{i1}, \ldots, G_{ip}$ are the marginal distributions of F_i.

Assume that C_i are continuous.

Copula $\equiv$ CDF with uniform margins
$\sim$ scaling the margins to expose the dependence structure.

Empirical Copula

For Population $i \in \{1, \ldots, m\}$ and $\ell \in \{1, \ldots, p\}$ fixed,
let $R_{ij\ell}$ be the ranks of the data $X_{ij\ell}$.

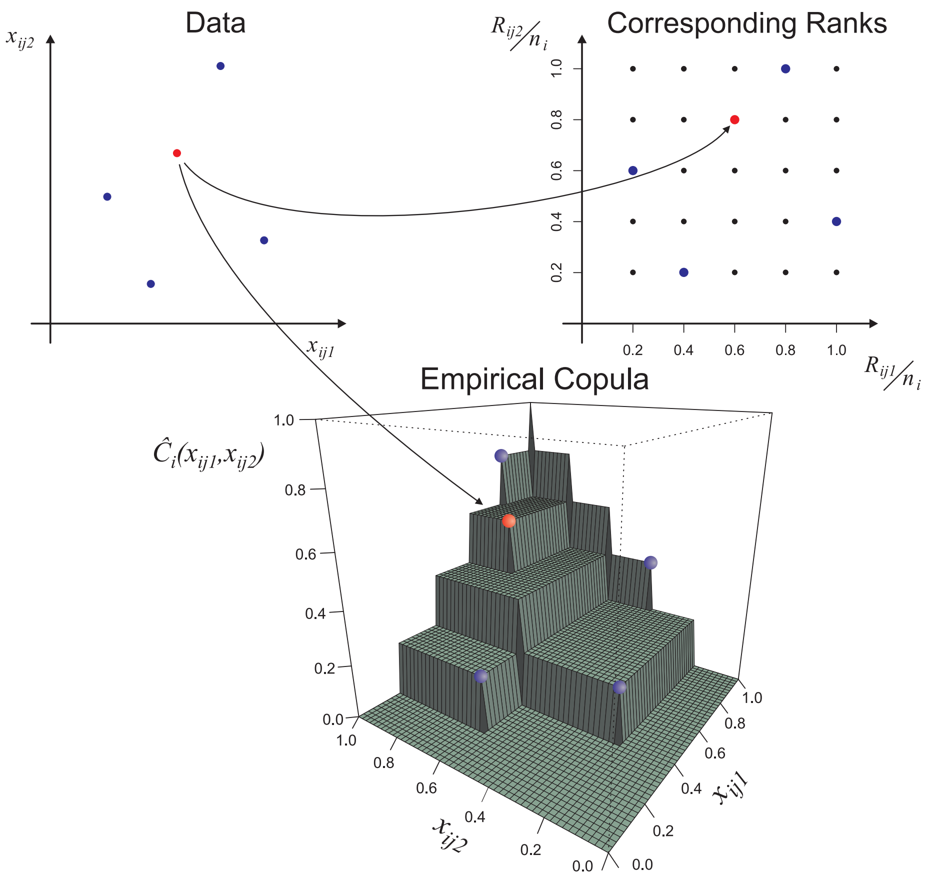
The empirical copula

$$\hat{C}_i(u_1, \ldots, u_p) = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbb{1}\left(\frac{R_{ij1}}{n_i} \leq u_1, \ldots, \frac{R_{ijp}}{n_i} \leq u_p \right)$$

allocates a mass of $1/n_i$ to each point

$$\left(\frac{R_{ij1}}{n_i}, \ldots, \frac{R_{ijp}}{n_i} \right).$$

Ranks are invariant to a monotone transformation of the margins...



MAMSE Weights

We choose the weights $\lambda_i \geq 0$ with $\sum_{i=1}^{m} \lambda_i = 1$ minimizing

$$P_{\boldsymbol{\lambda}} = \int |\hat{C}_1(\mathbf{u}) - \hat{C}_{\boldsymbol{\lambda}}(\mathbf{u})|^2 + \widehat{\mathrm{var}}\{\hat{C}_{\boldsymbol{\lambda}}(\mathbf{u})\} \,\mathrm{d}M(\mathbf{u})$$

where $\widehat{\mathrm{var}}\{\hat{C}_{\boldsymbol{\lambda}}(\mathbf{u})\} = \sum_{i=1}^{m} \frac{\lambda_i^2}{n_i} \hat{C}_i(\mathbf{u})\{1 - \hat{C}_i(\mathbf{u})\}$ is an approximation.

The measure $\mathrm{d}M$ allocates an equal mass of $1/n_1^p$
to each of the p-dimensional points

$$\left\{\frac{1}{n_1}, \frac{2}{n_1}, \ldots, 1\right\} \times \cdots \times \left\{\frac{1}{n_1}, \frac{2}{n_1}, \ldots, 1\right\}.$$

Weighted Pseudo-Likelihood

The family of copulas $C(\mathbf{u}|\theta)$, admitting densities $c(\mathbf{u}|\theta)$, is used to model the data. The value of θ maximizing

$$L(\theta) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} c\left(Y_{ij1}, \ldots, Y_{ijp}|\theta\right)^{\lambda_i/n_i}$$

is called the maximum weighted pseudo-likelihood estimate (MWPLE).

The Y_{ijp} are ranks scaled to $(0,1)$. Typically, $\quad Y_{ijp} = \dfrac{R_{ijp}}{n_i + 1}$

Suppose: a) sample sizes go to ∞,
b) λ_i are the MAMSE weights.

Weighted Empirical Copula

$$\hat{\mathcal{C}}(\mathbf{u}) = \sum_{i=1}^{m} \lambda_i \hat{C}_i(\mathbf{u})$$

is such that

$$\sup_{\mathbf{u}\in[0,1]^p} |\hat{\mathcal{C}}(\mathbf{u}) - C_1(\mathbf{u})| \to 0 \quad a.s.$$

Maximum Weighted Pseudo-Likelihood Estimate

If the parameter space is compact, the MWPLE based on MAMSE weights is strongly consistent.

Simulation

Measurement Error in Multiple Dimensions

The target population is a multivariate normal with covariance matrix

$$\Sigma_A = \begin{bmatrix} 1 & 0.4 & 0.3 & 0.2 \\ 0.4 & 1 & 0.4 & 0.3 \\ 0.3 & 0.4 & 1 & 0.4 \\ 0.2 & 0.3 & 0.4 & 1 \end{bmatrix} \text{ or } \Sigma_B = \begin{bmatrix} 1 & 0.8 & 0.6 & 0.4 \\ 0.8 & 1 & 0.8 & 0.6 \\ 0.6 & 0.8 & 1 & 0.8 \\ 0.4 & 0.6 & 0.8 & 1 \end{bmatrix}.$$

Samples from four populations of 4-dimensional data are generated. Population 1 is clean, but populations 2, 3 and 4 have measurement errors that affect their dependence structure.

		$100\times$			
	n	$\bar{\lambda}_1$	$\bar{\lambda}_2$	$\bar{\lambda}_3$	$\bar{\lambda}_4$
Scenario A	20	46	18	18	18
Scenario B	20	41	20	20	19

TABLE 5. Average weights for 4D data with measurement error.

		100 MSE(MPLE)/MSE(MWPLE)						
	n	$\mathbf{\Gamma}_1$	Γ_{11}	Γ_{12}	Γ_{13}	Γ_{14}	Γ_{15}	Γ_{16}
Scenario A	20	235	232	234	259	225	234	229
Scenario B	20	98	58	118	214	59	130	62

TABLE 6. Efficiency for 4D data with measurement error.

Note that $\mathbf{\Gamma}_1 = [\Gamma_{11}, \Gamma_{12}, \Gamma_{13}, \Gamma_{14}, \Gamma_{15}, \Gamma_{16}]^\mathsf{T}$
is the vector of correlations in the covariance matrix for Population 1.