Recent Statistical Advances

Instructions: Click on the link to access each author's presentation.

Chair: Andrés Christen García

Participants:

Zhu Wang: Unified Robust Estimation

Eman Aboaldahab Elsayed*

Charles Iwuji*

Awoke Seyoum Tegegne: Prevalence for the Disclosure of HIV Status to Sexual Partners and Its Determinants among Adults under cART in Amhara Region, Northwest Ethiopia

* *Work presentation not available or non-existent*

Unified Robust Estimation

Zhu Wang

University of Tennessee Health Science Center
USA

IAOS-ISI 2024 Mexico

May 2024

Loss function in regression

- The model function has the form

$$f(x, \beta) = \beta_0 + \beta_1 x \tag{1}$$

- The model function has the form

$$f(x,\beta) = \beta_0 + \beta_1 x \tag{1}$$

-

$$\underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - f(x_i,\beta))^2 \tag{2}$$

Loss function in regression

- The model function has the form

$$f(x, \beta) = \beta_0 + \beta_1 x \tag{1}$$

- $$\underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - f(x_i, \beta))^2 \tag{2}$$

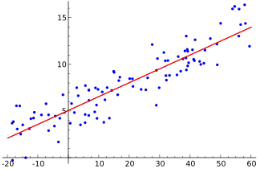

Figure 1 Linear regression

Data are available on the log of the surface temperature and the log of the light intensity of 47 stars in the star cluster CYG OB1.

Figure 2 Star temperature

Estimation in the presence of outliers

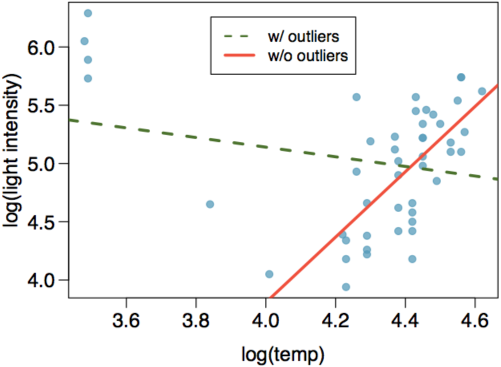

Figure 3 Linear regression

Robust linear regression

- Huber, Andrews or Tukey's biweight loss

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i}^{n} \Gamma(y_i, \mathbf{x}_i^\top \boldsymbol{\beta}), \tag{3}$$

- iteratively reweighted least squares (IRLS):

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i}^{n} w_i (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 \tag{4}$$

Challenges of robust estimation

- High-dimensional data (variable selection)
- Nonlinear relationship and high-order interactions (machine learning)
- Nonconvex loss (reliable algorithm)
- Weighted estimation and extensions to GLM (unified framework)

- robust loss functions: composite of concave and convex functions, called CC-family.
- iteratively reweighted convex optimization (IRCO)

Robust loss functions

Table 1 Composite loss functions with $\sigma > 0$ unless otherwise specified.

Type	Loss function $g(s(u))$	$g(z)$	$s(u)$
Regression			
Huber	$\begin{cases} \frac{u^2}{2} & \text{if } \lvert u\rvert \leq \sigma, \\ \sigma\lvert u\rvert - \frac{\sigma^2}{2} & \text{if } \lvert u\rvert > \sigma. \end{cases}$	$\begin{cases} z & \text{if } z \leq \sigma^2/2, \\ \sigma(2z)^{\frac{1}{2}} - \frac{\sigma^2}{2} & \text{if } z > \sigma^2/2. \end{cases}$	$\frac{u^2}{2}$
Andrews	$\begin{cases} \sigma(1-\cos(\frac{u}{\sigma})) \\ \quad \text{if } \lvert u\rvert \leq \sigma\pi, \\ 2\sigma \text{ if } \lvert u\rvert > \sigma\pi. \end{cases}$	$\begin{cases} \sigma(1-\cos(\frac{(2z)^{\frac{1}{2}}}{\sigma})) \\ \quad \text{if } z \leq \sigma^2\pi^2/2, \\ 2\sigma \text{ if } z > \sigma^2\pi^2/2. \end{cases}$	$\frac{u^2}{2}$
Biweight	$1-(1-\frac{u^2}{\sigma^2})^3 I(\lvert u\rvert \leq \sigma)$	$1-(1-\frac{2z}{\sigma^2})^3 I(z \leq \sigma^2/2)$	$\frac{u^2}{2}$
ClossR	$1-\exp(\frac{-u^2}{2\sigma^2})$	$1-\exp(\frac{-z}{\sigma^2})$	$\frac{u^2}{2}$
Classification			
Closs	$1-\exp(\frac{-(1-u)^2}{2\sigma^2})$	$1-\exp(\frac{-z}{\sigma^2})$	$\frac{(1-u)^2}{2}$
Rhinge	$1-\exp(-\frac{\max(0,1-u)}{2\sigma^2})$	$1-\exp(\frac{-z}{2\sigma^2})$	$\max(0, 1-u)$
Thinge	$\min(1-\sigma, \max(0, 1-u))$, $\sigma \leq 0$	$\min(1-\sigma, z)$	$\max(0, 1-u)$
Tlogit	$\min(1-\sigma, \log(1+\exp(-u)))$, $\sigma \leq 0$	$\min(1-\sigma, z)$	$\log(1+\exp(-u))$
Texp	$\min(1-\sigma, \exp(-u))$, $\sigma \leq 0$	$\min(1-\sigma, z)$	$\exp(-u)$
Dlogit	$\log(1+\exp(-u)) - \log(1+\exp(-u-\sigma))$	$\log(\frac{1+z}{1+z\exp(-\sigma)})$	$\exp(-u)$
Gloss	$\frac{1}{(1+\exp(au))^\sigma}$, $\sigma \geq 1, a > 0$	$(\frac{z}{1+z})^\sigma$	$\exp(-au)$
Qloss	$1-\int_{\infty}^{\frac{u}{\sigma}} \frac{1}{\sqrt{2\pi}} \exp(\frac{-x^2}{2})dx$	$1-\frac{1}{\sqrt{\pi}}\int_0^{\frac{z}{\sigma^2}} \frac{\exp(-t)}{\sqrt{t}}dt$	$\frac{u^2}{2}$

Concave convex family

Unified robust loss functions (Wang, 2024)
Let g : range $s \to \mathbb{R}$, $s : \mathbb{R} \to \mathbb{R}$, and the domain of g and s is a convex set. Concave convex (CC) family contains functions Γ satisfying the following conditions:

i. $\Gamma = g \circ s$
ii. g is a nondecreasing closed concave function on range s
iii. $\partial(-g(z))\ \forall z \in$ range s is nonempty and bounded
iv. s is convex on $\mathbb{R}$.

We call g concave component, and s convex component. Members of the CC-family are nonconvex if g is bounded.

Subgradient and subdifferential

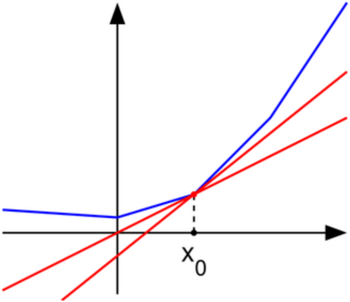

Figure 4 A convex function (blue) and "subtangent lines" at x_0 (red).
https://en.wikipedia.org/wiki/Subderivative

Table 2 Concave component with $\sigma > 0$.

Concave	$g(z), z \geq 0$	Source
hcave	$\begin{cases} z & \text{if } z \leq \sigma^2/2, \\ \sigma(2z)^{\frac{1}{2}} - \frac{\sigma^2}{2} & \text{if } z > \sigma^2/2. \end{cases}$	Huber
acave	$\begin{cases} \sigma^2(1 - \cos(\frac{(2z)^{\frac{1}{2}}}{\sigma})) & \text{if } z \leq \sigma^2\pi^2/2, \\ 2\sigma^2 & \text{if } z > \sigma^2\pi^2/2. \end{cases}$	Andrews
bcave	$\frac{\sigma^2}{6}\left(1 - (1 - \frac{2z}{\sigma^2})^3 I(z \leq \sigma^2/2)\right)$	Biweight
ccave	$\sigma^2\left(1 - \exp(\frac{-z}{\sigma^2})\right)$	Closs
dcave	$\frac{1}{1-\exp(-\sigma)} \log(\frac{1+z}{1+z\exp(-\sigma)})$	Dlogit
ecave	$\begin{cases} \frac{2\exp(-\frac{\delta}{\sigma})}{\sqrt{\pi\sigma\delta}} z & \text{if } z \leq \delta, \\ \operatorname{erf}(\sqrt{\frac{z}{\sigma}}) - \operatorname{erf}(\sqrt{\frac{\delta}{\sigma}}) + \frac{2\exp(-\frac{\delta}{\sigma})}{\sqrt{\pi\sigma\delta}}\delta & \text{if } z > \delta. \end{cases}$	Qloss
gcave	$\begin{cases} \frac{\delta^{\sigma-1}}{(1+\delta)^{\sigma+1}} z & \text{if } z \leq \delta, \\ \frac{1}{\sigma}(\frac{z}{1+z})^\sigma - \frac{1}{\sigma}(\frac{\delta}{1+\delta})^\sigma + \frac{\delta^\sigma}{(1+\delta)^{\sigma+1}} & \text{if } z > \delta. \end{cases}$ where $\delta = \begin{cases} \to 0+ & \text{if } 0 < \sigma < 1, \\ \frac{\sigma-1}{2} & \text{if } \sigma \geq 1. \end{cases}$	Gloss
tcave	$\min(\sigma, z), \sigma \geq 1$ for classification; $\sigma > 0$ otherwise	Truncation

Table 3 Convex component.

Convex	$s(u)$
Gaussian	$\frac{u^2}{2}$
GaussianC	$\frac{(1-u)^2}{2}$
Binomial	$\log(1+\exp(-u))$
Exponential family	$-\left(\frac{yu-b(u)}{a(\phi)}+c(y,\phi)\right)$
Hinge	$\max(0, 1-u)$
ϵ-insensitive	$\begin{cases} 0 & \text{if } \lvert u\rvert \leq \epsilon, \\ \lvert u\rvert-\epsilon & \text{if } \lvert u\rvert > \epsilon. \end{cases}$

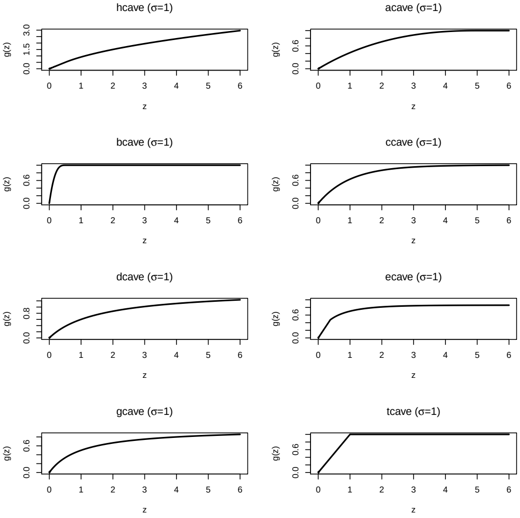

Figure 5 Concave component.

Operations that preserve CC-family

Theorem (Nonnegative weighted sums)

Let $\Gamma_1 = g_1 \circ s$ *and* $\Gamma_2 = g_2 \circ s$ *be members of the CC- family* Ω *and* $c_1, c_2 \geq 0, g = c_1 g_1 + c_2 g_2$*. Then* $\Gamma = g \circ s \in \Omega$ *holds and*

$$\partial(-g(z)) = c_1 \partial(-g_1(z)) + c_2 \partial(-g_2(z)) \tag{5}$$

for any z *from int (dom* g*) = int (dom* g_1*)* $\cap$ *int (dom* g_2*), where int (dom* g*) is the interior of domain of* g*.*

Theorem (Minimization)

Let $\Gamma_i = g_i \circ s, i = 1, ..., m,$ *be members of the CC-family* $\Omega, g = \min_{1 \leq i \leq m} g_i$*. Then* $\Gamma = g \circ s \in \Omega$ *holds and for any* $z \in$ *int (dom* g*)* $= \cap_{i=1}^{m}$*int (dom* g_i*) we have*

$$\partial(-g(z)) = Conv\{\partial(-g_i(z)) | i \in I(z)\}, \tag{6}$$

$$Conv\{x_1, ..., x_m\} = \{x = \sum_{i=1}^{m} a_i x_i | a_i \geq 0, \sum_{i=1}^{m} a_i = 1\},$$

$$I(z) = \{i : g_i(z) = g(z)\}.$$

Robust loss for regression

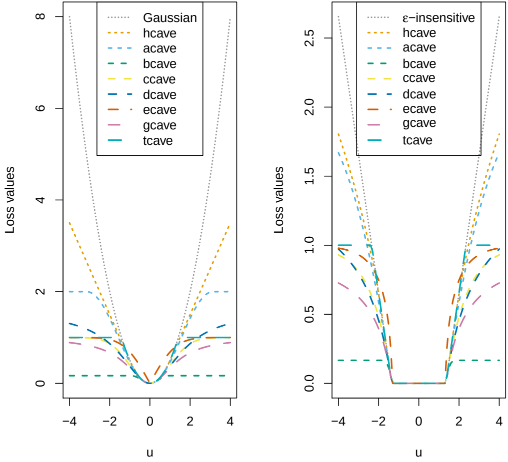

Figure 6 Convex component Gaussian, ϵ-insensitive and their induced composite loss functions.

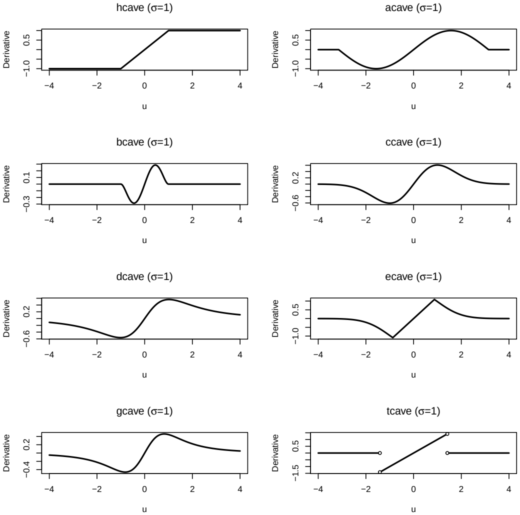

Figure 7 Derivatives of Gaussian induced composite loss functions

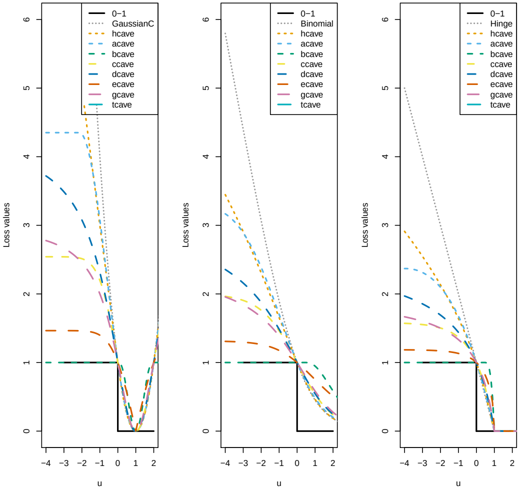

Figure 8 Convex component GaussianC, Binomial, Hinge loss and their induced composite loss functions.

Fisher consistency

1. $s(u) < s(-u), \;\; u > 0$.
2. $s'(0) < 0$.
3. g : range of $s \rightarrow \mathbb{R}$ is strictly increasing.
4. $g'(s(0)) \neq 0$ exists.
5. $g \circ s$ is a non-increasing function with $\sigma \geq 1$.
6. If $\sigma = 1$, then $1 = g(s(0)) > g(s(1))$ and $g(s(0)) = g(s(-1))$ hold.
7. If $\sigma > 1$, then $g'(s(0)) \neq 0$ exists.

Theorem

Assume that $\Gamma = g \circ s$. Then for $Y \in \{-1, 1\}, \Gamma(Yf(X))$ is Fisher-consistent if either of the following two sets of conditions holds:

(i) *Conditions 1–4 hold.*

(ii) *Conditions 2, 5–7 hold.*

- We have data $(x_{ij}, y_i), i = 1, ..., n, j = 0, 1, ..., p$, where x_{ij} is the predictor and y_i is the response variable. Let $\mathbf{x}_i = (x_{i0}, ..., x_{ip})^\mathsf{T}$ denote a $(p+1)$-dimensional predictor with the first entry 1, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^\mathsf{T}$ a $(p+1)$-dimensional coefficient vector and β_0 the intercept.

- We have data $(x_{ij}, y_i), i = 1, \dots, n, j = 0, 1, \dots, p$, where x_{ij} is the predictor and y_i is the response variable. Let $\mathbf{x}_i = (x_{i0}, \dots, x_{ip})^\top$ denote a $(p+1)$-dimensional predictor with the first entry 1, $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^\top$ a $(p+1)$-dimensional coefficient vector and β_0 the intercept.
- Consider convex component $s(u_i)$ given in Table 3, where u_i is linked to the linear predictor $f_i = \mathbf{x_i}^\top \boldsymbol{\beta}$:

$$u_i = \begin{cases} y_i - f_i, & \text{for regression,} \\ y_i f_i, & \text{for classification with } y_i \in [-1, 1], \\ f_i, & \text{for exponential family.} \end{cases} \tag{7}$$

- We have data $(x_{ij}, y_i), i = 1, ..., n, j = 0, 1, ..., p$, where x_{ij} is the predictor and y_i is the response variable. Let $\mathbf{x}_i = (x_{i0}, ..., x_{ip})^{\mathsf{T}}$ denote a $(p+1)$-dimensional predictor with the first entry 1, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^{\mathsf{T}}$ a $(p+1)$-dimensional coefficient vector and β_0 the intercept.
- Consider convex component $s(u_i)$ given in Table 3, where u_i is linked to the linear predictor $f_i = \mathbf{x_i}^{\mathsf{T}}\boldsymbol{\beta}$:

$$u_i = \begin{cases} y_i - f_i, & \text{for regression,} \\ y_i f_i, & \text{for classification with } y_i \in [-1, 1], \\ f_i, & \text{for exponential family.} \end{cases} \tag{7}$$

- A CC-estimator is a solution that minimizes the empirical loss $L(\boldsymbol{\beta})$ given by

$$L(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} L_i(\boldsymbol{\beta}), L_i(\boldsymbol{\beta}) = g(s(u_i)), \tag{8}$$

where g and s are the concave and convex component in the CC-family, respectively.

- In many applications, we optimize a penalized loss function $F : \mathbb{R}^{p+1} \to \mathbb{R}$:

$$F(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) + \Lambda(\boldsymbol{\beta}), \tag{9}$$

where

$$\Lambda(\boldsymbol{\beta}) = \sum_{j=1}^{p} \left(\alpha p_\lambda(|\beta_j|) + \lambda \frac{1-\alpha}{2} \beta_j^2 \right),$$

$0 \leq \alpha \leq 1, \lambda \geq 0$, and $p_\lambda(|\beta_j|)$ is the penalty function such as the LASSO (Tibshirani, 1996) or SCAD (Fan and Li, 2001).

Robust penalized estimation

- In many applications, we optimize a penalized loss function $F : \mathbb{R}^{p+1} \to \mathbb{R}$:
$$F(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) + \Lambda(\boldsymbol{\beta}), \tag{9}$$
where
$$\Lambda(\boldsymbol{\beta}) = \sum_{j=1}^{p} \left(\alpha p_\lambda(|\beta_j|) + \lambda \frac{1-\alpha}{2} \beta_j^2 \right),$$
$0 \leq \alpha \leq 1, \lambda \geq 0$, and $p_\lambda(|\beta_j|)$ is the penalty function such as the LASSO (Tibshirani, 1996) or SCAD (Fan and Li, 2001).
- Minimizing the penalized loss function can avoid overfitting, provide shrinkage estimates and conduct variable selection. The loss function in (8) is a special case of (9) with $\Lambda(\boldsymbol{\beta}) = 0$, i.e., $\lambda = 0$.

Algorithm design by the first-order condition of convexity

- g is concave

$$g(u) \leq g(\hat{u}) + g'(\hat{u})(u - \hat{u}). \tag{10}$$

- can define a surrogate function $\gamma(u|\hat{u})$

- the convex or Fenchel conjugate function of $h(z)$:

$$\varphi(v) = \sup_{z \in \mathrm{dom}\ h} (zv - h(z)).$$

- And conjugate of $\varphi(v)$ is restored if $h(z)$ is a closed convex function (Lange, 2016, Fenchel-Moreau theorem):
- can define another surrogate function

Main result

Algorithm 1 IRCO Algorithm

1: **Initialize** $\beta^{(0)}$ and set $k = 0$
2: **repeat**
3: Compute $u_i(\beta^{(k)})$ in (7) and $z_i = s(u_i(\beta^{(k)})), i = 1, \ldots, n$
4: Compute $v_i^{(k+1)}$ via $v_i^{(k+1)} \in \partial(-g(z_i))$ or $z_i \in \partial\varphi(v_i^{(k+1)}), i = 1, \ldots, n$
5: Compute $\beta^{(k+1)} = \text{argmin}_\beta \sum_{i=1}^n s(u_i(\beta))(-v_i^{(k+1)}) + \Lambda(\beta)$
6: $k = k + 1$
7: **until** convergence of $\beta^{(k)}$

Concave	$\partial(-g(z))$
hcave	$\begin{cases} -1 & \text{if } z \leq \sigma^2/2 \\ -\sigma(2z)^{-\frac{1}{2}} & \text{if } z > \sigma^2/2 \end{cases}$
acave	$\begin{cases} -\frac{\sigma \sin(\frac{\sqrt{2z}}{\sigma})}{\sqrt{2z}} & \text{if } 0 < z \leq \sigma^2\pi^2/2 \\ -1 & \text{if } z = 0 \\ 0 & \text{if } z > \sigma^2\pi^2/2 \end{cases}$
bcave	$-\frac{1}{\sigma^4}(2z-\sigma^2)^2 I(z \leq \sigma^2/2)$
ccave	$-\exp(-\frac{z}{\sigma^2})$
dcave	$-\frac{\exp(\sigma)}{(z+1)(z+\exp(\sigma))}$
ecave	$\begin{cases} -\frac{2}{\sqrt{\pi\sigma\delta}} \exp(\frac{-\delta}{\sigma}) & \text{if } z \leq \delta \\ -\frac{2}{\sqrt{\pi\sigma z}} \exp(\frac{-z}{\sigma}) & \text{if } z > \delta \end{cases}$
gcave	$\begin{cases} -\frac{\delta^{\sigma-1}}{(\delta+1)^{\sigma+1}} & \text{if } z \leq \delta \\ -\frac{z^{\sigma-1}}{(z+1)^{\sigma+1}} & \text{if } z > \delta \end{cases}$
tcave	$\begin{cases} \{-1\} & \text{if } z < \sigma \\ \{0\} & \text{if } z > \sigma \\ [-1, 0] & \text{if } z = \sigma \end{cases}$

Table 4 Subdifferential of negative concave component.

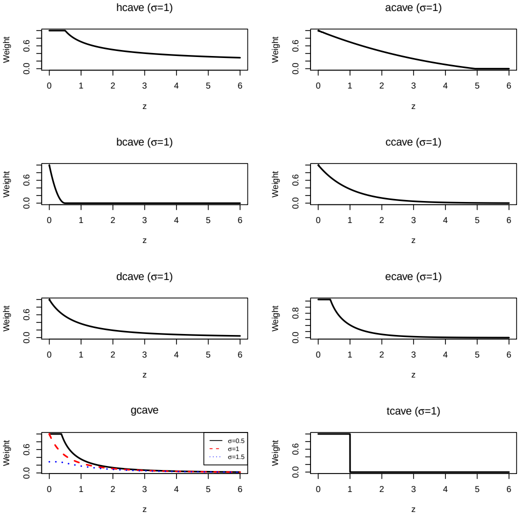

Figure 9 Weight function $-\partial(-g(z))$.

Theorem

Suppose that g is a concave component in the CC-family, and g is bounded below.

1. *The loss function values $F(\boldsymbol{\beta}^{(k)})$ generated by Algorithm 1 are nonincreasing and converge.*
2. *Assume that g and s are differentiable, $\zeta(u, v) = s(u)(-v) + \varphi(v)$ is jointly continuous in (u, v), φ is the conjugate function of $-g$, $\nabla L(\boldsymbol{\beta}) = \nabla \ell(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)})$, where the surrogate loss is given by*

$$\ell(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)}) = \sum_{i=1}^{n} \zeta(u(\boldsymbol{\beta}), v(\boldsymbol{\beta}^{(k)})),$$

and $p_\lambda(|\cdot|)$ satisfies mild assumptions. Then every limit point of the iterates generated by Algorithm 1 is a Dini stationary point of $F(\boldsymbol{\beta})$.

Applications

- Robust (penalized) least squares in regression
- Robust (penalized) least squares in classification
- Robust (penalized) generalized linear models
- Robust support vector machine
- Robust support vector machine in regression
- R package `mpath` https://cran.r-project.org/package=mpath

- In a UK hospital, 135 expectant mothers were surveyed on the decision of human milk feeding their babies or not, along with two-level predictive factors (Heritier et al., 2009).

Robust logistic regression

- In a UK hospital, 135 expectant mothers were surveyed on the decision of human milk feeding their babies or not, along with two-level predictive factors (Heritier et al., 2009).
- We compute binomial-induced CC-estimators, i.e., robust logistic regression, and display the robust weights in Figure 10. The subjects 3, 11, 14, 53, 63, 75, 90 and 115 have smallest weights, confirming a more complex estimator in Heritier et al. (2009).

Robust logistic regression

- In a UK hospital, 135 expectant mothers were surveyed on the decision of human milk feeding their babies or not, along with two-level predictive factors (Heritier et al., 2009).
- We compute binomial-induced CC-estimators, i.e., robust logistic regression, and display the robust weights in Figure 10. The subjects 3, 11, 14, 53, 63, 75, 90 and 115 have smallest weights, confirming a more complex estimator in Heritier et al. (2009).
- For variable selection, we develop a usual SCAD logistic regression

Robust logistic regression

- In a UK hospital, 135 expectant mothers were surveyed on the decision of human milk feeding their babies or not, along with two-level predictive factors (Heritier et al., 2009).
- We compute binomial-induced CC-estimators, i.e., robust logistic regression, and display the robust weights in Figure 10. The subjects 3, 11, 14, 53, 63, 75, 90 and 115 have smallest weights, confirming a more complex estimator in Heritier et al. (2009).
- For variable selection, we develop a usual SCAD logistic regression
- This λ value is then utilized to compute binomial-induced SCAD CC-estimators. The estimated coefficients of the selected variables are shown in Table 5.

Robust logistic regression

- In a UK hospital, 135 expectant mothers were surveyed on the decision of human milk feeding their babies or not, along with two-level predictive factors (Heritier et al., 2009).
- We compute binomial-induced CC-estimators, i.e., robust logistic regression, and display the robust weights in Figure 10. The subjects 3, 11, 14, 53, 63, 75, 90 and 115 have smallest weights, confirming a more complex estimator in Heritier et al. (2009).
- For variable selection, we develop a usual SCAD logistic regression
- This λ value is then utilized to compute binomial-induced SCAD CC-estimators. The estimated coefficients of the selected variables are shown in Table 5.
- The CC-estimators provide coefficients of `smokenowYes` < -2. Being a smoker during pregnancy has larger negative effect from robust estimation.

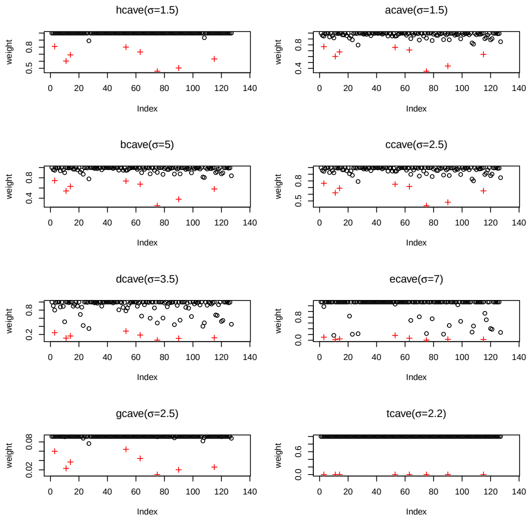

Figure 10 Robustness weights of logistic regression for the human milk feeding data.

Table 5 Estimates of robust penalized logistic regression for the breastfeeding data.

Variable	logis	hcave	acave	bcave	ccave	dcave	ecave	gcave	tcave
(Intercept)	0.10	−0.20	0.32	0.33	0.35	2.71	3.27	−0.70	−2.27
pregnancyBeginning									
howfedBreast						0.12			
howfedfrBreast	1.05	1.42	1.19	1.21	1.18	0.03	0.05	1.76	1.27
partnerPartner	0.48	0.24	0.20	0.13	0.22				
smokenowYes	−2.00	−2.31	−2.38	−2.44	−2.38	−3.89	−4.25	−2.69	−2.48
smokebfYes									
age									
educat		0.03	0.01	0.01	0.01			0.06	0.16
ethnicNon-white	1.94	2.49	2.52	2.64	2.48	1.16	2.45	3.25	3.59

Summary

- CC-estimation is an iteratively reweighted estimation procedure for robust estimation, powerful for nonconvex problems
- Zhu Wang (2024), Unified Robust Estimation, http://dx.doi.org/10.1111/anzs.12409
- R package `mpath` https://cran.r-project.org/package=mpath
- Zhu Wang (2021), Unified Robust Boosting, https://arxiv.org/abs/2101.07718
- R package `irboost` https://cran.r-project.org/package=irboost

- NIH/NIDDK R21DK130006

References I

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Heritier, S., Cantoni, E., Copt, S., and Victoria-Feser, M.-P. (2009). *Robust Methods in Biostatistics*, volume 825. Chichester, England: John Wiley & Sons.

Lange, K. (2016). *MM Optimization Algorithms*. Philadelphia: SIAM.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

Wang, Z. (2024). Unified Robust Estimation. *Australian & New Zealand Journal of Statistics*, 66(1):77–102.

PREVALENCE FOR THE DISCLOSURE OF HIV STATUS TO SEXUAL PARTNERS AND ITS DETERMINANTS AMONG ADULTS UNDER CART IN AMHARA REGION, NORTH-WEST ETHIOPIA

By Awoke Seyoum Tegegne(Prof.)
Bahir Dar University
Ethiopia

IAOS-ISI 2024
MEXICO CONFERENCE

INEGI
International Statistical Institute
IAOS
IMPROVING OFFICIAL STATISTICS

Outline of the presentation

- Introduction
- Materials and participants
- Results
- Discussion
- Conclusion and recommendation

INTRODUCTION

- Currently, HIV/AIDS continues to be a serious global public health problem.
- It is the cause of 36.7 million people living with HIV and 1.8 million new infections each year. •
- The problem is also the cause of one million people dying from HIV-related cases .
- Among these, about 19.4 million people are testified to live with HIV in Eastern and Southern Africa .
- In Ethiopia, the problem seems to be stable given that it is different in different regions in the country.
- According to the Ethiopian Public Health Institute (EPHI) report, the Amhara Region, one of the eleven regions in the country, accounts for the highest number of people living with HIV.
- In the region, the overall incidence rate of new HIV infection is 6.9 per 1000 tested population.
- Several factors are responsible for reducing the infection, which can be grouped as economic, social, and demographic factors

Introduction…

- One way of reducing the spread of the disease may be **e**ncouraging people living with the virus to disclose their disease status to their sexual partners.
- This is may be important to reduce the transmission of HIV by making awareness and decreasing risky behavior.
- Disclosure of the HIV status facilitates other behaviors that may improve the management of HIV.
- Previous studies indicate that individuals who disclosed their HIV diagnosis results have better adherence to ART treatments .
- Female adults who disclose their HIV status to sexual partners are more likely to participate in the prevention of mother-to-child HIV transmission programs.

INTRODUCTION…

- Studies previously conducted indicate that disclosure may increase opportunities to receive social support, which may help individuals cope and recover from physical illness, and decrease depressive symptoms due to HIV-related indications .
- Disclosure of HIV-positive status to all societies living around them is crucial for HIV avoidance and provision of care.
- Hence, it is important to discover the prevalence of disclosure of HIV status to sexual partners and its factors determining individuals not disclose their HIV positive status in order to reduce the transmission of the disease to uninfected people.

Introduction...

- Disclosure of HIV-positive status to all societies living around them is crucial for HIV avoidance and provision of care.
- Hence, it is important to discover the prevalence of disclosure of HIV status to sexual partners and its factors determining individuals not disclose their HIV positive status in order to reduce the transmission of the disease to uninfected people.
- Among studies conducted previously in the developed world, rates of the disclosure of HIV disease to sexual partners ranged from 42% up to 100%, depending on the large part on the type of partner to whom the person disclosed.
- The previous studies also indicate that the rate of disclosure of the disease in developing countries is lower than the rates reported in developed countries.
- The rates of disclosure in developing countries vary from 16.7% to 86% with average disclosure of 49%.

INTRODUCTION…

- To the best of our knowledge, there is limited region-wide research on the prevalence of disclosure of HIV status to sexual partners and its predictors among HIV positive adults under cART.
- The issue of disclosure of HIV status increases opportunities for implementation of HIV risk reduction, improving access to treatment, and motivating partners for Voluntary Counseling and Testing (VCT) activities .Thus, disclosure of HIV status is an issue to be addressed for HIV prevention and treatment.
- *The objective of this study was to determine the prevalence of disclosure of HIV status to sexual partners and its associated factors among adults living with HIV/AIDS (PLWAs) in the Amhara Region, Ethiopia.*
- *The other objective of the current study was to check whether the results obtained in developed country also true in developing country.*
- *The result obtained in the current investigation is important for regional policy makers to make evidence-based HIV prevention and interventions.*

MATERIALS AND PARTICIPANTS

Study Area and Population

- *The* study was conducted in the Amhara Region (northwest Ethiopia).
- The region is one of the nine well-known regions in the country with a large population ,the second next to the Oromia region.
- The region has 12 zones, three-city administrations, and 180 woredas(139 rural and 41 urban).
- According to the Ethiopian Central Statistics Agency, the region has a projected population of 21.5 million people, of which 80% of them are rural farmers.
- The region has only 80 public hospitals, 847 health centers, and 3,342 health posts.
- Amhara's healthcare system is unable to modernize and provide quality health services due to many challenges particularly; the transmission rate of the disease from one individual to another is still a series of problems. This is why the region was selected as a study area.. The study population in the current investigation was all HIV-positive adults under treatment

STUDY DESIGN

- A retrospective cohort study design was conducted on 792 randomly selected adult HIV-infected patients under cART in the Amhara Region, Northwest Ethiopia.
- In the hospital, about 6 thousand people with HIV were receiving treatment and of these, about 2 thousand of them were under cART.
- The data were taken in ART sections of Felege Hiwot Teaching and Specialized Hospital and its catchment areas.
- The hospital is a specialized, teaching, and referral with a regional laboratory, where all HIV patients throughout the region are referred to this hospital and all treatment results are sent to this hospital for a regional laboratory experiment. Finally, the regional laboratory results are organized and sent to the Federal Ministry of Health.

SOURCE OF DATA

- Secondary data collected from participants' charts by the health staff for treatment purposes were used in the current investigation.

Participants.

- The source populations for the current investigation were all HIV-positive adults under cART and following their treatment at zonal and district hospitals and treatment results were sent to Felege Hiwot Teaching and Specialized Hospital, Amhara Region, Ethiopia.
- The study population was adults who fulfilled the inclusion criteria.

Sample Size and Sampling Procedures.

- Random samples of 792 HIV-positive adults were selected considering their ART identification number.
- Cochran's formula is used for calculating the sample size when the population is large. Cochran (1977) developed a formula to calculate a representative sample for proportions

- *HIV-Infected* Patients under cART with at least two visits to the treatment site whose follow-ups were from January 2015 up to December 2020 were included in this investigation.

Variables under Current Investigation

- ***Th**e* dependent variable for this study was disclosure of the HIV status to sexual partners among HIV-positive adults under cART. It has two levels namely disclosed and not disclosed.
- The disease is said to be disclosed if a sexual partner had full information about the status of the disease, otherwise, it is not disclosed.
- Since the patients considered under this investigation are under cART (combined antiretroviral conducted correctly.
- The predictor variables were sex, age, marital status, level of education, social support, social violence, residential area,
- the existence of mental depression, religion, functional status, opportunistic infectious disease, WHO stages of HIV, adherence levels, and baseline CD4 cell count.

SELF-REPORTED PREDICTOR VARIABLES

- Other predictors, like dietary instruction, the time when pills were taken, the existence of mental depression, the existence of social violence by people living together, the existence of social support, and the existence of medication allergic at the initial time were reported by participants and recorded carefully in each patient's chart.
- The reason for monthly follow-ups at the initial time was to follow up on whether there existed medication side effects like mental depression, skin scratch, and any other medication allergic on individuals at the initial time.

STATISTICAL MODEL AND DATA ANALYSIS IN THE CURRENT STUDY

- In this investigation, an analysis of binary data in terms of the binomial distributions with logit transformation was conducted. The result is a binomial response conducted with a logistic regression model with a logit link function.
- Descriptive statistics were conducted to assess basic participants' characteristics.
- Bivariate analysis was conducted to determine the presence of statistically significant correlations between explanatory variables and the outcome variable.
- A binary logistic regression model was used for investigating the variable of interest.
- Model selection was assessed using the stepwise selection technique.
- Odds ratios (OR) and their 95% CI were also used to look into the significant effect between the dependent and independent variables.

SELF REPORTED ADHERENCE LEVEL

- In this investigation, a person was categorized as food adherent if he/she always followed dietary instructions directed by the health staff, otherwise, he/she was categorized as non-adherent .
- Similarly, a patient was categorized as time adherent if he/she always followed time scheduling instructions given by the health practitioners otherwise categorized as non-adherent.
- Patients' self-report on whether drug medication had been skipped or not were used to assess adherence to medication. Based on this, a person was said to be non-adherent to medication, if he/she took <95% of the prescribed pills.
- If a patient's adherence is ≥95% of the prescribed medication, he/she is categorized as adherent to medication.

IOS-ISI 2024
MEXICO CONFERENCE

DATA COLLECTION TOOLS AND QUALITY OF DATA

- The data collection tools/format were developed by the investigator in consultation with the health staff at the ART section of the hospital and the quality of data was controlled by the health staff at the ART section.
- To assure the quality of the data, the questionnaire was pretested on PLWHA (5% of the sample size i.e., 40 individuals) and amendments were incorporated to the questionnaire to obtain full information on the variables included in the investigation.
- Statistical Analysis System (SAS) version 9.4software was used to analyze the data.
- A binary logistic regression model was employed for the longitudinal outcome variable (disclosure of the HIV status of their sexual partner). A statistical decision was made at a 5% level of significance.
- The goodness of fit for the current investigation was conducted using the Akaike information criterion (AIC) and Bayesian information criterion (BIC), considering the model with the smallest AIC and BIC as the best of all others.

RESULTS

- As shown in Table 1, out of the sample of 792 patients, 40.9% were rural residents, 50.6% were females, 56.3% were living with their partners, 21% disclosed their disease to family members, and 49.2% were owners of cell phones.
- Only 25.5% of the patients were adherent and the rest were non-adherent.
- Finally, among the respondents, less than one-third of the patients disclosed their disease status to sexual partners (21%).
- Among the participants who disclosed their HIV status,17.3% disclosed the disease status on the day of receiving the test result, 18.5% disclosed their status within a week, 9.7% of them disclosed their disease status within 2 weeks, and the remaining of them disclosed their disease status within a month

RESULT…

- Reasons for non-disclosure of the disease status were recorded by the health staff and some of the reasons were:35% as fear of separation/divorce, 37.7% of them said that their partner might be afraid of the transmission of HIV from them, 25.5% of the other said fear of accusation of disloyalty, 7.1% of the participants not disclose because of fear of being labeled as a bad person, 5% of them said that no enough time to discuss because their partner works in other place, and 6.1% declared that because of fear of physical abuse.
- As shown in Table 1, out of the sample of 792 patients, 40.9% were rural residents, 50.6% were females, 56.3% were living with their partners, 21% disclosed their disease to family members, and 49.2% were owners of cell phones.
- Only 25.5% of the patients were adherent and the rest were non-adherent. Finally, among the respondents, more than 50% of them (79%) did not disclose the disease to sexual partners.

TABLE 1: *Baseline sociodemographic and clinical variables of 792 participants in the study area.*

Variable		Average (Q1, Q3)	No (%)
Weight (kg)		58.1 (45–70)	—
Baseline CD4 cells/mm^3		148.7 (113–180)	—
Age (years)		64.3 (48–78)	—
Follow-up times		23 visits	—
First month/initial CD4 cell count change/mm^3		16.6 (12–26)	—
Sex	Male		392 (49.4)
	Female		400 (50.6)
Educational status	No education		163 (20.6)
	Primary		209 (26.4)
	Secondary		274 (34.6)
	Tertiary		146 (18.4)
Residential area	Urban		468 (59.1)
	Rural		324 (40.9)
Existence of social violence	Yes		345 (56.4)
	No		447 (43.6)
Existence of mental depression	Yes		478 (60.4)
	No		314 (39.6)
Availability of social support	Yes		350 (44.2)
	No		442 (55.8)
Marital status	Living with partner		446 (56.3)
	Living without partner		346 (43.7)
WHO HIV stage	Stage I		101 (12.8)
	Stage II		259 (32.7)
	Stage III		199 (25.1)
	Stage IV		233 (29.4)
Disclosure of the disease	Yes		166 (21.0)
	No		226 (79.0)
Adherence to cART	Adherent		202 (25.5)
	Non-adherent		590 (74.5)

RESULTS...

- Among the participants who disclosed their HIV status, 17.3% disclosed the disease status on the day of receiving the test result, 18.5% disclosed their status within a week, 9.7% of them disclosed their disease status within 2 weeks, and the remaining of them disclosed their disease status within a month.
- *Reasons for non-disclosure of the disease status were recorded by the health staff and some of the reasons were;* 35% as fear of separation/divorce**,** *25.5% of the other said fear of accusation of disloyalty, 7.1% of the participants not disclose because of fear of being labeled as a bad person, 5% of them said that no enough time to discuss because my partner works in other place and 6.1% declared that because of fear of physical abuse*

RESULT…

- Parameter estimation which helps to identify statistically significant predictors for the variable of interest is indicated in Table 2.
- Table 2 indicates that predictors like age of patients, baseline CD4 cell count, the number of followed-up visits, marital status, sex, residential area, opportunistic infectious diseases, level of education, and level of adherence to cART had a significant effect on the variable of interest.
- The result in Table 2, revealed that Age of patients, the number of visits by the patients, patients with good cART adherence, female patients, patients living with their partner, Patients living in urban area, educated HIV patients had positive effect for the disclosure of HIV disease for sexual partners.
- On the other hand, existence of social violence, existence of opportunistic disease, non-educated patients, Patients with WHO early stages(Stage I) negatively affected for the disclosure of HIV disease status for sexual partners.

TABLE2: PARAMETER ESTIMATES FOR DISCLOSURE OF HIV STATUS FOR SEXUAL PARTNERS

Parameter	Estimates	St. error	Adjusted odds ratio (AOR)	Wald 95% CI		*p*-value
Intercept	3.01	0.03	20.287	11.53	58.62	<0.001*
Age	0.02	0.01	1.020	1.001	1.120	0.004*
Baseline CD4 cell count	−0.02	0.01	0.980	0.764	0.991	<0.001*
Follow-up times	0.01	0.01	1.010	1.002	1.034	<0.001*
Marital status (Ref. = Without partner)						
With partners	0.01	0.021	1.010	1.003	1.112	0.006*
Sex (Ref. = Male)						
Female	0.01	0.012	1.010	1.001	1.021	0.007*
Residential area (Ref. = Urban)						
Rural	−0.02	0.023	0.980	0.96	0.99	0.004*
Level of education (Ref. = educated)						
Non-educated	−0.05	0.452	0.950	0.92	0.98	0.003*
Adherence (Ref. = adherent)						
Non-adherent	−0.06	0.471	0.940	0.71	0.97	<0.001*
Existence of social violence (Refer = yes)						
No	0.012	0.354	1.012	1.008	1.234	<0.001*
Opportunistic infectious disease (Ref. = yes)						
No	0.06	0.521	1.062	1.049	1.191	0.002*
WHO stages (Ref. = stage IV)						
Stage I	−0.12	0.347	0.887	0.645	0.921	0.013*
Stage II	−0.13	0.065	0.878	1.12	1.05	0.021*
Stage III	−0.10	0.048	0.905	0.09	1.10	0.010*

*stands for significant variables at 95% CI.

Discussion

- The prevalence of this study indicates that among the total participants, only 21% of them disclosed their HIV status to their sexual partners.
- This indicates that the prevalence was very low as compared to the average rate of prevalence conducted in other developing countries(49%). The potential reason for this might be cultural, social, and economic factors.
- Potential predictors have been identified for different levels of disclosure of the disease status as discussed below. This needs further study.
- Overall, the level of disclosure of HIV-positive results in this study is below the rate of disclosure status at developing countries (49%). This indicates that more health-related educational work is needed to rise up the disclosure level of the HIV disease.
- Different patients disclosed the status of the disease at different times and only 17.3% disclosed the disease status on the day of receiving the test result.

CONCLUSION…

Among the predictors,

- Age of patients, follow-up visits, living with partners, female patients, non-existence of social violence, non-opportunistic disease, and being educated patients were positively associated with the increase of disclosure of the HIV disease status.

whereas

- the existence of social violence, being non-adherent to cART, non-educated patients, male patients, living without partner, and baseline CD4 cell count were negatively associated with disclosure of HIV disease status.

RECOMMENDATION

- Health-related education for HIV-positive adults to disclose their HIV status is a crucial issue.
- Knowledge of HIV transmission is also important to reduce the violence and discrimination against those HIV positive adults who disclosed their disease status.
- Special support for that HIV-infected individual who disclosed the disease may encourage the others to disclose their disease status without fear and anxiety.

Thank you for attention!!!