Optimal probability weights for inference with constrained precision

RSS 2016 Conference - Manchester

Michele Santacatterina
michele.santacatterina@ki.se

Matteo Bottai
matteo.bottai@ki.se

Unit of Biostatistics, Karolinska Institutet

September 7th 2016

Outline

```
> head(hmohiv)
  ID time age drug censor    entdate     enddate
1  1    5  46    0      1 5/15/1990  10/14/1990
2  2    6  35    1      0 9/19/1989   3/20/1990
3  3    8  30    1      1 4/21/1991  12/20/1991
4  4    3  30    1      1  1/3/1991    4/4/1991
5  5   22  36    0      1 9/18/1989   7/19/1991
6  6    1  32    1      0 3/18/1991   4/17/1991

> m1 <- coxph(Surv(time,censor)~drug,data=hmohiv)
> coef(summary(m1))[3]
[1] 0.2418138

> mw <- coxph(Surv(time,censor)~drug,data=hmohiv, weights = w)
> coef(summary(mw))[3]
[1] 8.123295

> summary(w)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.00    0.00    0.00    1.00    0.00   79.34
```

Objective

1. We have target probability weights, w^*.
2. The weighted estimate has large variance.
3. We estimate the weights closest to w^* within a variance constraint.

We propose a general method to estimate optimal probability weights based on the solution of a nonlinear constrained optimization problem.

Introduction

In statistics, probability weights are used in many areas of research including

- complex survey designs,
- missing data analysis,
- adjustment for confounding factors, etc.

Methods have been proposed to alleviate the sometimes excessive imprecision of weighted inference [1, 2, 3, among others]. In medical sciences the most frequent approach is weight trimming, or truncation, which consists of replacing outlying weights with less extreme ones.

Optimal probability weights

Let $\hat{\theta}_{w^*}$ be an unbiased estimator for a population parameter θ^* that uses weights $w^* = (w_1^*, \ldots, w_n^*)^T$, with $\mathbf{1}^T w^* = 1$ and $w^* \geq 0$. Let σ_{w^*} indicate the standard error of $\hat{\theta}_{w^*}$ and $\hat{\sigma}_{w^*}$ an estimator for it. Instead of trimming the weights, we suggest deriving the weights $\hat{w}$ that are closest to w^* with respect to the Euclidean norm $\|w - w^*\|$, under the constraint that the estimated standard error $\hat{\sigma}_{\hat{w}}$ be less than or equal to a specified constant $\xi > 0$.

$$
\begin{aligned}
\underset{w \in \mathbb{R}^n}{\text{minimize}} \quad & \|w - w^*\| && (1) \\
\text{subject to} \quad & \hat{\sigma}_w \leq \xi && (2) \\
& \mathbf{1}^T w = 1 && (3) \\
& w \geq 0 && (4)
\end{aligned}
$$

When a solution $\hat{w}$ to problem (1)-(4) exists, constraint (2) guarantees that the estimated standard error of the estimator with weights $\hat{w}$ is less than or equal to ξ. Constraints (3) and (4) guarantee that the optimal weights $\hat{w}$ are bounded and non-negative, respectively.

Properties

(i) *Consistency.* The probability that $\hat{\theta}_{\hat{w}} = \hat{\theta}_{w^*}$ converges to one if $\hat{\sigma}_w$, the estimator for the standard error for the weighted estimator, converges to zero as the sample size tends to infinity, for any set of probability weights $\hat{\mathbf{w}}$ and any constant value ξ.

(ii) *Minimum-bias estimator.* The optimally-weighted estimator $\hat{\theta}_{\hat{w}}$, obtained using $\hat{\mathbf{w}}$, is the the estimator with minimum bias among all weighted estimators with standard error less or equal than ξ.

(iii) *Uniqueness.* If the nonlinear constrained optimization problem is convex, then the set of optimal weights $\hat{\mathbf{w}}$ is unique. In this case, by property (i) and (ii), the optimally-weighted estimator is the unique minimum-bias estimator among all weighted estimators with constrained precision.

Lagrange multiplier

The Lagrange multiplier λ in constraint (6) and the value of the objective function at the optimum can be used to choose the level of precision ξ. More specifically, large values of λ suggest that minimal changes in ξ would cause large changes in the objective function. Large values of the objective function at the optimum indicate that the set of optimal weights are far from the target set.

$$
\begin{aligned}
\underset{w \in \mathbb{R}^n}{\text{minimize}} \quad & \|w - w^*\| && (5) \\
\text{subject to} \quad & \hat{\sigma}_w \leq \xi && (6) \\
& \mathbf{1}^T w = 1 && (7) \\
& w \geq 0 && (8)
\end{aligned}
$$

Case study

We evaluated the effect of early initiation on time to virological failure across subgroups. We used data from the Swedish InfCare HIV registry.

Four known factors for HIV-treatment progression were considered:

1 logarithm of viral load, ln(VL), at treatment initiation,

2 age at treatment initiation,

3 route of transmission, and

4 gender.

Table: Subgroups considered for the analysis of the optimal timing of HIV treatment initiation.

Subgroup	ln(VL)	Age	Route	Gender
1	10.5	31	IDU	Female
2	10.5	31	IDU	Male
3	10.5	31	Hetero	Female
4	10.5	31	Hetero	Male
5	10.5	31	MSM	Male
6	10.5	31	Other	Female
7	10.5	31	Other	Male
8	10.5	46	IDU	Female
9	10.5	46	IDU	Male
10	10.5	46	Hetero	Female
11	10.5	46	Hetero	Male
12	10.5	46	MSM	Male
13	10.5	46	Other	Female
14	10.5	46	Other	Male

Early initiation was defined as HIV-treatment initiation with 500+ CD4 cells/μ. Virological failure happens when the treatment fails to suppress the HIV virus.

Target populations

We defined the target populations $f_j(x)$, $j = 1, \ldots, 14$,

$$f_j(x) = \begin{cases} \phi(\ln(\mathrm{VL}) - 10.5)\,\phi(age - \mu_j) & \text{if } x = \left(\ln(\mathrm{VL}), \mathrm{age}, \mathrm{route}_j, \mathrm{gender}_j\right) \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

where $x = (\ln(\mathrm{VL}), \mathrm{age}, \mathrm{route}, \mathrm{gender})$, and ϕ is the standard normal distribution. Standard deviations were set equal to 1.

Subgroup (j)	ln(VL)	Age (μ_j)	Route	Gender
1	10.5	31	IDU	Female
2	10.5	31	IDU	Male
3	10.5	31	Hetero	Female
4	10.5	31	Hetero	Male
5	10.5	31	MSM	Male
6	10.5	31	Other	Female
7	10.5	31	Other	Male
8	10.5	46	IDU	Female
9	10.5	46	IDU	Male
10	10.5	46	Hetero	Female
11	10.5	46	Hetero	Male
12	10.5	46	MSM	Male
13	10.5	46	Other	Female
14	10.5	46	Other	Male

Optimal weights

Target weights were calculated as

$$\hat{w}_j^* = f_j(x) / \hat{f}_0(x), \tag{10}$$

where $\hat{f}_0(x)$ is the multivariate density kernel estimate for ln(VL), age, route of transmission and gender in the sampled population. For each target population, we computed the optimal probability weights $\hat{\mathbf{w}}$ by solving the nonlinear constrained problem, where $\hat{\sigma}_w$ denotes the estimated standard error of the estimator for the parameter β_w in

$$\lambda_i(t) = \lambda_0(t) \exp\left(\beta_w I\left[CD4_{i,0\in(500+)}\right]\right), \tag{11}$$

$i = 1, \ldots, n$. The indicator function $I\left[CD4_{i,0\in(500+)}\right]$ is equal to 1 if individuals started treatment with CD4 cell count above 500 cells/μL, and 0 otherwise. We evaluated values for the Lagrange multiplier λ and the objective function over a range of different values for ξ starting from high precision, $\xi = 1$, to the precision of the unweighted estimator, $\xi = \hat{\sigma}_{\beta,w^*}$, when the constraint is inactive.

Subgroup	$\lambda/1000$	$\sqrt{n}\|\|w-w^*\|\|$	β_{w^*}	$\sigma^2_{w^*}$	$\beta_{\hat{w}}$	$\sigma^2_{\hat{w}}$	ξ
1	0.03	0.38	6.8	5e+02	3.37	10	10
2	106.38	34.39	−14.82	1e+06	−3.27	20	20
3	0	0	0.39	7e−02	0.39	0.07	1
4	0	0.01	6.23	1e+00	6.22	1	1
5	0	0	1.9	4e−02	1.9	0.04	10
6	26.14	15.72	9.83	1e+03	5.2	10	10
7	0.77	0.62	−10.2	2e+02	−5.16	1	1
8	120.38	60.61	0.9	2e+06	0.05	30	30
9	0	0	1.23	4e−02	1.23	0.04	1
10	4.28	7.66	−11.46	3e+02	−8.47	15	15
11	0	0.09	−9.51	4e+02	−5.36	5	5
12	0	0	0.49	2e−02	0.49	0.02	10
13	102.36	31.8	1.36	1e+06	2.02	10	10
14	120.64	60.59	0.3	2e+06	0.05	30	30
Overall	0	0	0.3	2e−02	0.3	0.02	0.02

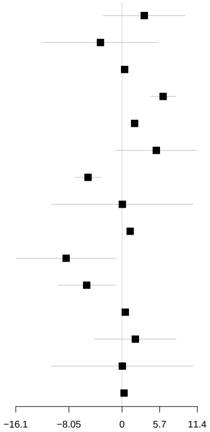

Figure: *Optimal timing of HIV treatment initiation across subgroups.* Lagrange multiplier in (2), square root of the objective function, target-weighted coefficient $\hat{\beta}_{w*}$, variance for $\hat{\beta}_{w*}$, optimally-weighted coefficient $\hat{\beta}_{\hat{w}}$, variance for $\hat{\beta}_{\hat{w}}$, and chosen level ξ.

Conclusions

- Probability weights are used in many settings;
- The variance of weighted estimators can be large;
- The proposed method can estimate the probability weights closest to the target weights within a variance constraint.

References

[1] F. Potter, “A study of procedures to identify and trim extreme sampling weights,” in *Proceedings of the American Statistical Association, Section on Survey Research Methods*, vol. 225230, 1990.

[2] J.-F. Beaumont, “A New Approach to Weighting and Inference in Sample Surveys,” *Biometrika*, vol. 95, pp. 539–553, Sept. 2008.

[3] D. Pfeffermann, “Inference under informative sampling,” in *Sample Surveys: Inference and Analysis* (D. Pfeffermann and C. R. Rao, eds.), pp. 455–487, Elsevier, Oct. 2009.

[4] Q. Li and J. Racine, “Li, Q. and Racine, J.S: Nonparametric Econometrics: Theory and Practice. (eBook and Hardcover),” 2007.

[5] J. Aitchison and C. G. G. Aitken, “Multivariate binary discrimination by the kernel method,” *Biometrika*, vol. 63, no. 3, pp. 413–420, 1976.

[6] Q. Li and J. Racine, “Nonparametric estimation of distributions with categorical and continuous data,” *Journal of Multivariate Analysis*, vol. 86, no. 2, pp. 266 – 292, 2003.

[7] T. Hayfield and J. Racine, “Nonparametric econometrics: The np package,” *Journal of Statistical Software*, vol. 27, no. 1, pp. 1–32, 2008.

[8] A. Wächter and L. T. Biegler, “On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming,” *Mathematical Programming*, vol. 106, pp. 25–57, Apr. 2005.

[9] HSL, “”HSL. A collection of Fortran codes for large scale scientific computation. ”,” 2016.

A1: Target weights estimation

We calculated the set of target weights as

$$\hat{w}_j^* = f_j\left(x\right)/\hat{f}_0\left(x\right). \tag{12}$$

We used generalized product kernels [4] to estimate $\hat{f}_0\left(x\right)$. The generalized product kernel function for the vector x, is the product of each kernel function, where continuous variables use the second order Gaussian kernel function, and discrete variables use the discrete kernel function suggested by [5]. We used the data-driven method of bandwidth selection for the generalized product kernels estimator developed by [6]. The R package "np" [7] was used in the analyses.

A2: Optimization algorithm

We solved the nonlinear constrained mathematical optimization problems with a primal-dual interior point algorithm.
Specifically, the R interface of Ipopt [8], "Ipoptr", was used. "Ipoptr" solves general large-scale nonlinear constrained optimization problems. The MA57 sparse symmetric system [9] was used as a line-search method within Ipopt.

A3: Simulations

In each scenario we randomly generated 1,000 samples each of which comprised 100 observations from a normally-distributed variable under the following model: $y_i \sim N(20 + 4x_i, 5)$, where $i = 1, \ldots, 100$, and $x_i \sim beta(x_i \mid \alpha_0, \beta_0)$, a beta distribution with parameters α_0 and β_0.
The target weights were defined as

$$w_i^* = \frac{beta(x_i \mid \alpha_1, \beta_1)}{beta(x_i \mid \alpha_0, \beta_0)}. \tag{13}$$

We considered fifty different scenarios, constructed by combining the following parameter values: $\alpha_0 = \{1, 2, 3, 4, 5\}$, $\beta_0 = \{1, 2, 3, 4, 5\}$, and $(\alpha_1, \beta_1) = \{(2, 5), (5, 5)\}$.
We considered two estimators for the weighted mean:

- the optimal estimator $\hat{\theta}_{\hat{w}} = y^T \hat{w}$,
- the trimmed estimator $\hat{\theta}_{\overline{w}} = y^T \overline{w}$.

Simulations

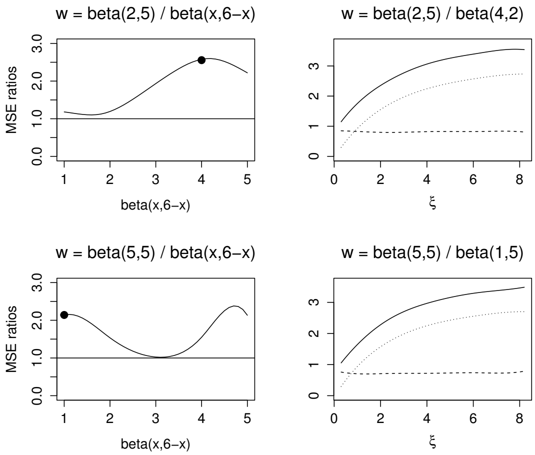

Figure: Left-hand-side panels: mean squared error ratio between trimmed and optimally weighted estimators. Right-hand-side panels: mean squared error (solid line), variance (dotted), and bias (dashed) of the optimally weighted estimator $\hat{\theta}_{\hat{w}}$, for different values of ξ in the scenarios.

A4: Proof of Property (ii) *Minimum-bias estimator.*

Suppose that the target weighted estimator, $\hat{\theta}_{w^*}$, is the solution to the weighted equation

$$\sum_{i=1}^{n} w_i^* h_i(\hat{\theta}_{w^*}) = 0, \tag{14}$$

where h_i is a known function of the sample data and the parameter θ. Applying a Taylor series expansion of $h_i(\hat{\theta}_{\hat{w}})$ around $\hat{\theta}_{w^*}$, it can be shown that the optimally-weighted estimator is the solution to

$$\sum_{i=1}^{n} \hat{w}_i \left[h_i(\hat{\theta}_{w^*}) + h_i'(\hat{\theta}_{w^*})(\hat{\theta}_{\hat{w}} - \hat{\theta}_{w^*}) + O((\hat{\theta}_{\hat{w}} - \hat{\theta}_{w^*})^2) \right] = 0. \tag{15}$$

A2: Proof of Property (ii) *Minimum-bias estimator.*

From equation 15, considering that the remainder O converges quadratically to zero as $(\hat{\theta}_{\hat{w}} - \hat{\theta}_{w^*})$ tends to zero, and that $E(\hat{\theta}_{w^*}) = \theta^*$, the bias of the optimally-weighted estimator is shown to be approximately equal to

$$E(\hat{\theta}_{\hat{w}} - \theta^*) = E(\hat{\theta}_{\hat{w}} - \hat{\theta}_{w^*}) + E(\hat{\theta}_{w^*}) - \theta^* \approx -E\left[\frac{(\hat{w} - w^*)^T h(\hat{\theta}_{w^*})}{\hat{w}^T \nabla_w h(\hat{\theta}_{w^*})}\right], \tag{16}$$

where $\nabla_w h(\hat{\theta}_{w^*})$ is the gradient of the vector $(h_1(\hat{\theta}_{w^*}), \ldots, h_n(\hat{\theta}_{w^*}))^T$. The optimally-weighted estimator is approximately unbiased for θ^* if the vectors $(\hat{w} - w^*)$ and $h(\hat{\theta}_{w^*})$ are orthogonal. Finally, by property (i), minimizing the objective function $\|\mathbf{w} - \mathbf{w}^*\|$ is equivalent to minimizing the bias of the optimally-weighted estimator with respect to the target parameter θ^*, yielding the minimum-bias estimator among all weighted estimators with precisions less or equal than ξ.