Institut de statistique

Une interprétation de la pseudo-vraisemblance - An interpretation of the pseudo-likelihood

Monique Graf
Institut de Statistique - Université de Neuchâtel, Switzerland
and Elpacos Statistics, la Neuveville, Switzerland

10e Colloque Francophone sur les Sondages
Université de Lyon

Outline

Introduction

One stage design

Two stage design

Discussion

Introduction

- Multilevel models= special case of the generalized mixed model, used for the analysis of survey data with several levels (strata, clusters, units)
- Binder (1983), Gourieroux et al.(1984), Skinner et al. (1989), Pfeffermann et al. (1998) : pseudo-likelihood for surveys with unequal inclusion probabilities.
- In multi-stage surveys, scaling of weights influence the parameter estimates (see e.g. Rabe-Hesketh and Skrondal, 2006 and Asparouhov, 2006).
- No theory on the choice of scaling.

Alternatives to the pseudo-likelihood

- Rao et al. (2013) propose a method by estimating functions that have good asymptotic properties.
- Sampling density conditional on the distribution of weights for non-ignorable designs, e.g. Pfeffermann (2011).
  Bonnéry et al. (2018) establish asymptotic properties of the likelihood obtained with this density.

Goal

- Given a postulated population distribution,
- obtain the pseudo-likelihood,
- find a proper likelihood
  - belonging to the same family of distributions as the population distribution
  - as "close" as possible to the pseudo-likelihood.
- Derive a method for rationally choosing the scaling of weights.

One stage

Consider a one stage design. Let

- $\{y_i, w_i, i = 1, ...n\}$ = sampled units and the corresponding extrapolation weights.
- y_i : realization of a random variable Y_i
- a model : Y_i are i.i.d with pdf $f(.;\boldsymbol{\theta})$ depending on a set of parameters $\boldsymbol{\theta}$.

Pseudo-log-likelihood

In a one-stage design, the pseudo-log-likelihood given by

$$\ell^{pseudo}(\boldsymbol{\theta}; \mathbf{y}, \mathbf{w}) = \sum_{i=1}^{n} w_i \log f(y_i, \boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(y_i, \boldsymbol{\theta})^{w_i}.$$

ℓ^{pseudo} is a proper log-likelihood, if it can be written as a sum of log-densities, up to a constant term not depending on the parameters.

1. Conditions for ℓ^{pseudo} to be a proper log-likelihood, ℓ^{proper} ?
2. Conditions for pdf $K_i^{-1} f(y_i, \boldsymbol{\theta})^{w_i}$ to belong to the same family of distributions as $f(y_i, \boldsymbol{\theta})$?
3. Conditions for the parameters of ℓ^{pseudo} and ℓ^{proper} to coincide ?

Correction term - One stage design

In general,

$$\int_{-\infty}^{\infty} f(y,\boldsymbol{\theta})^{xw_i}\, dy = K(xw_i,\boldsymbol{\theta}) = K_i \Longrightarrow K_i^{-1} f(y,\boldsymbol{\theta})^{xw_i} \text{ is a pdf.}$$

$$\ell^{proper} \doteq \sum_{i=1}^{n} \log[K(xw_i,\boldsymbol{\theta})^{-1} f(y,\boldsymbol{\theta})^{xw_i}]$$

Thus

$$\begin{aligned}
\ell^{pseudo} &= \ell^{proper} + \sum_i \log[K(xw_i,\boldsymbol{\theta})] - \sum_i xw_i \log[K(1,\boldsymbol{\theta})] \\
&= \ell^{proper} + C(x\mathbf{w},\boldsymbol{\theta}).
\end{aligned}$$

Equivalence condition

$$\ell^{pseudo} \text{ equivalent to } \ell^{proper}$$
$$\Longleftrightarrow$$
$$C(x\mathbf{w}, \boldsymbol{\theta}) = C(x\mathbf{w}).$$

Sampling pdf

- $K(w_i, \boldsymbol{\theta})^{-1} f(y, \boldsymbol{\theta})^{w_i}$ can be interpreted as the sampling pdf of Y_i, the random variable associated to the i-th sampled unit.
- observations are no longer identically distributed, but still independent (according to the model).
- the sampling pdf depends on the scaling of weights.

How to choose the scaling ?

Canonical scaling

- A proper likelihood is the sum of n log-densities where n is the sample size.
- $\tilde{w}_i, i = 1, ..., n =$ provided weights.
- *Canonical weights* :

$$w_i = n \frac{\tilde{w}_i}{\sum_{k=1}^{n} \tilde{w}_k} = \frac{\tilde{w}_i}{\bar{\tilde{w}}} \quad \text{sum to } n.$$

- Another scaling can always be defined from the canonical weights.
  $x =$ scaling factor
  $xw_i =$ scaled weight.

Normal distribution - One stage design

- $Y_i \sim N(\mu_i, \sigma^2)$
- $\mathbf{X}$ = matrix of auxiliary variables ;
- $\mathbf{x}_i^t$ = i-th row of $\mathbf{X}$
- $\mu_i = \mathrm{E}(Y_i) \doteq \mathbf{x}_i^t \boldsymbol{\beta}$
- parameters : $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma)$

Normal distribution - One stage design

- Population log-likelihood

$$\ell^{pop}(\boldsymbol{\beta}, \sigma; \mathbf{y}) = \sum_{i=1}^{N} \log \left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{(y_i - \mu_i)^2}{\sigma^2} \right) \right).$$

- Pseudo-log-likelihood

$$\ell^{pseudo}(\boldsymbol{\beta}, \sigma; \mathbf{y}, x\mathbf{w}) = \sum_{i=1}^{n} xw_i \log \left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{(y_i - \mu_i)^2}{\sigma^2} \right) \right]$$

- Proper log-likelihood

$$\ell^{proper}(\boldsymbol{\beta}, \sigma; \mathbf{y}, x\mathbf{w}) = \sum_{i=1}^{n} \log \left[\frac{\sqrt{xw_i}}{(\sigma\sqrt{2\pi})} \exp\left(-\frac{1}{2} \frac{(y_i - \mu_i)^2}{(\sigma/\sqrt{xw_i})^2} \right) \right]$$

- Correction term

$$C(x, \mathbf{w}, \boldsymbol{\theta}) = \sum_{i=1}^{n} (xw_i - 1) \log \left(\frac{1}{\sigma\sqrt{2\pi}} \right) - \frac{1}{2} \log \left(\prod_{i=1}^{n} xw_i \right)$$

Normal distribution - One stage design

- The correction term can be simplified,

$$C(x, \mathbf{w}, \sigma) = n \left\{ (x-1) \log \left(\frac{1}{\sigma\sqrt{2\pi}} \right) - \frac{1}{2} [\log(x) + \log(G)] \right\}$$

  where G is the geometric mean of the canonical weights.
- C does not depend on $\boldsymbol{\beta}$.
- $\hat{\boldsymbol{\beta}}^{pseudo}$ and $\hat{\sigma}^{pseudo}$ do not depend on x.
- $\ell^{proper}(\boldsymbol{\beta}, \sigma; \mathbf{y}, x\mathbf{w}) \equiv \ell^{proper}(\boldsymbol{\beta}, \sigma/\sqrt{x}; \mathbf{y}, \mathbf{w})$ thus

$$\hat{\sigma}^{pseudo} = \widehat{\sigma_x}^{proper} \text{ where } \sigma_x = \sigma/\sqrt{x}.$$

- C does not depend on σ if and only if $x = 1$.
  With the canonical weights, it is equivalent to estimate the parameters using the pseudo- or the proper log-likelihood.

Exponential distribution - One stage design

$$g(y;b) = \frac{1}{b}\exp\left(-\frac{y}{b}\right) \qquad y > 0;\ b > 0.$$

$$\begin{aligned} g^{w}(y;b) &= \left(\frac{1}{b}\right)^{w}\exp\left(-\frac{wy}{b}\right) = \frac{1}{b/w}\exp\left(-\frac{y}{b/w}\right)\,\frac{1}{wb^{w-1}} \\ &= g(y;b/w)\,\frac{1}{wb^{w-1}}. \end{aligned}$$

The pseudo-log-likelihood is given by

$$\ell^{pseudo}(b;\mathbf{y},x\mathbf{w}) = \sum_{i=1}^{n} xw_i \log\left(g(y_i;b)\right)$$

$$= \quad \ell^{proper}(b;\mathbf{y},x\mathbf{w}) - n\log(xG) - \sum_{i=1}^{n}(xw_i - 1)\log(b).$$

$$C(x,\mathbf{w},b) = -n\left\{\log(x) + (x-1)\log(b) + \log(G)\right\}$$

Same form as before.

Generalized gamma distribution - One stage design

- Probability density of $Y \sim GG(a, b, p)$:

$$g(y; a, b, p) \;=\; \frac{a}{\Gamma(p)}(y/b)^{ap} \exp\{-(y/b)^a\}\frac{1}{y} \quad a, b, p > 0.$$

  In the applications, $b = \exp(\mathbf{x}^t \beta)$, where $\mathbf{x}$ is a vector of auxiliary variables.

- Change of variable : $u = \log(y)$; pdf of $\log(Y)$:

$$f(u; a, b, p) \;=\; \frac{a}{\Gamma(p)}(e^u/b)^{ap} \exp\{-(e^u/b)^a\}$$

Which pseudo-likelihood ?

$$\ell^{pseudo} \text{ based on } g \quad \neq \quad \ell^{pseudo} \text{ based on } f$$

- ℓ^{pseudo} based on g : weights are applied to $\mathbf{y}$
- ℓ^{pseudo} based on f : weights are applied to $\log(\mathbf{y})$

Weights do not have the same meaning according to the model.

Good reason to choose f :
the sampling density is more similar to the population density.

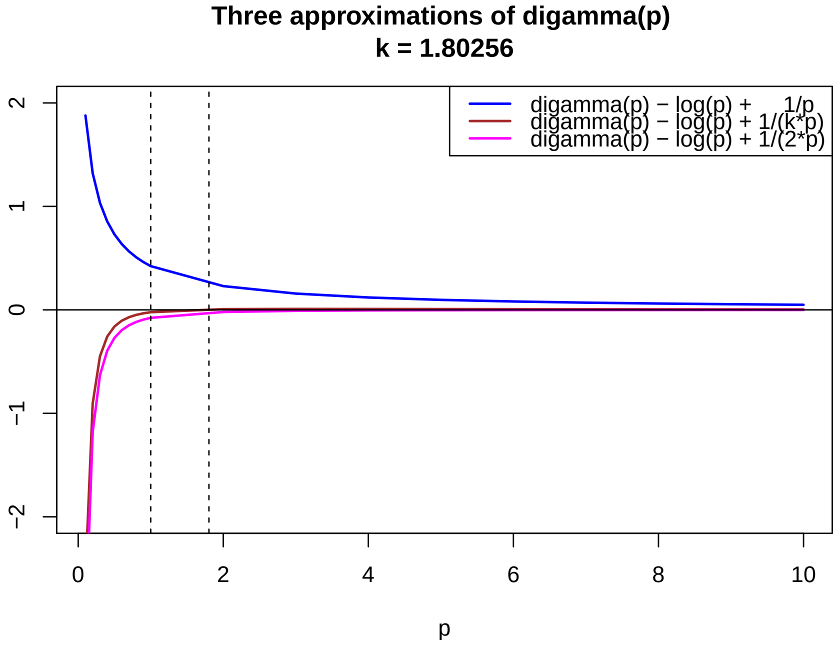
Generalized gamma distribution - One stage design

The proper log-likelihood is the sum of log-densities, pdf of $GG(a, b/(xw_i)^{1/a}, pxw_i)$.
Correction term for the pseudo-log-likelihood :

$$
\begin{aligned}
C(x, \mathbf{w}, a, p) &= \sum_i \log \left\{ \left[\frac{a}{\Gamma(p)} \right]^{xw_i} \frac{\Gamma(pxw_i)}{a} \right\} \\
&= n(x-1)\log(a) - nx\log(\Gamma(p)) + \sum_i \log(\Gamma(pxw_i))
\end{aligned}
$$

- C does not depend on b
- if $x = 1$, C does not depend on a
- if $w_i \neq 1$, the dependence on p remains.

With unequal weights, ℓ^{pseudo} and ℓ^{proper} will give different estimates.


Three approximations of digamma(p)
k = 1.80256
digamma(p) − log(p) + 1/p
digamma(p) − log(p) + 1/(k*p)
digamma(p) − log(p) + 1/(2*p)
−2
−1
0
1
2
0
2
4
6
8
10
p

Generalized gamma distribution - One stage design

Set $x = 1$.
$C_1(p) = C(1, \mathbf{w}, a, p) = -n \log(\Gamma(p)) + \sum_i \log(\Gamma(p\, w_i))$.

$$\frac{\partial}{\partial p}\ell^{pseudo} = \frac{\partial}{\partial p}\ell^{proper} + \frac{d}{dp}C_1(p),$$

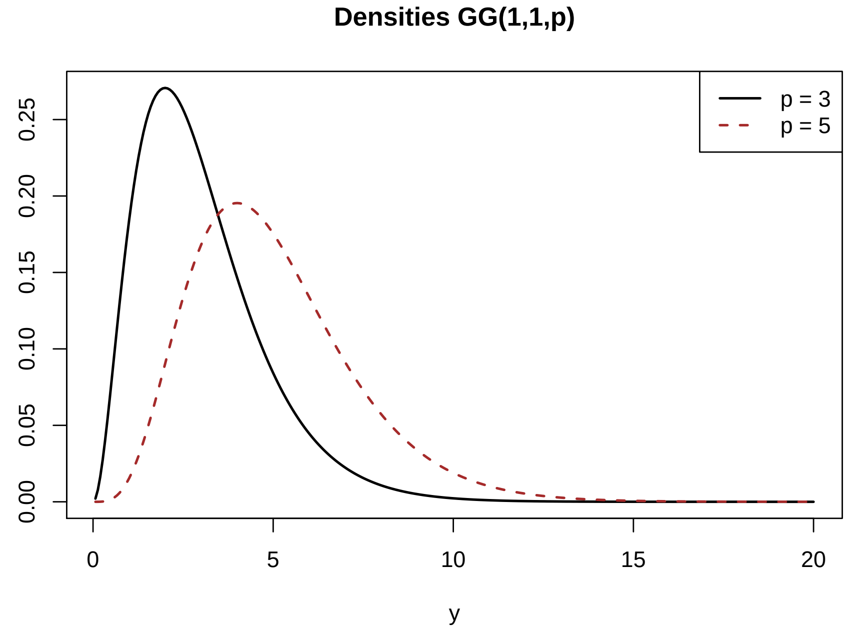
$$\begin{aligned}
\frac{d}{dp}C_1(p) &= -n\psi(p) + \sum_i w_i \psi(p\, w_i) \\
&\approx -n\left[\log(p) - \frac{1}{kp}\right] + \sum_i w_i \left[\log(w_i p) - \frac{1}{k w_i p}\right] \\
&= \sum_i w_i \log(w_i). \\
\frac{d}{dp}C_1(p) &= \sum_i w_i \log(w_i) \pm \frac{1}{2p}.
\end{aligned}$$

Generalized gamma distribution - One stage design

- Noufaily and Jones (2013) unweighted case : the score equation in p is strictly decreasing for given values of a and b.
- This property extends to the weighted case.
- It can be shown that if $n \geq 3$, $\sum_i w_i \log(w_i)$ is always positive.

Thus in general we expect

$$\hat{p}^{pseudo} > \hat{p}^{proper}.$$


Densities GG(1,1,p)
p = 3
p = 5
0.00
0.05
0.10
0.15
0.20
0.25
0
5
10
15
20
y

Two stage design

Primary sampling units (PSU) are selected and within each PSU a sample is selected.
Hypothesis : The model includes an additive random effect that corresponds to the PSU of the design.
Within each PSU j, weights $\tilde{w}_{ij}$ are provided for the ultimate unit $i, i \in j$.

- n_j = sample size in PSU j.
- $w_{ij} = n_j \frac{\tilde{w}_{ij}}{\sum_{k=1}^{n} \tilde{w}_{kj}} = \frac{\tilde{w}_{ij}}{\bar{\tilde{w}}_j}$ = canonical weight within primary unit j.
- observations within PSU j are conditionally independent given random effect V_j,
- $f_1(y - v; \boldsymbol{\theta})$ = conditional pdf of Y_{ij} given the random effect $V_j = v$.
- within PSU pseudo-log-likelihood = $\ell_j^{pseudo}(\boldsymbol{\theta}; \mathbf{y}_j - v\mathbf{1}_{n_j}, x\mathbf{w}_j) = \sum_{i=1}^{n_j} x w_{ij} \log[f_1(y_{ij} - v; \boldsymbol{\theta})]$

Two stage design

- V = latent unobserved PSU effect with pdf $f_2(v; \mathbf{\Theta})$.
- $\tilde{W}_j$ = provided weight of PSU $j, j = 1, ..., c$.
- W_j = canonical weight of PSU j,

$$W_j = \sum_{k=1}^{c} n_k \frac{\tilde{W}_j}{\sum_{k=1}^{c} n_k \tilde{W}_k} = \frac{\tilde{W}_j}{\bar{\tilde{W}}_n}.$$

Total sample size :

$$\sum_j n_j = \sum_j n_j W_j.$$

Two stage design

Total pseudo-log-likelihood =

$$
\begin{aligned}
& \ell^{pseudo}(\boldsymbol{\theta}, \boldsymbol{\Theta}; \{\mathbf{y}_j, x\mathbf{w}_j, j = 1, .., c\}; t\mathbf{W}) \\
= & \sum_{j=1}^{c} tW_j \log \left[\int_{-\infty}^{\infty} \exp(\ell_j^{pseudo}(\boldsymbol{\theta}; \mathbf{y}_j - v\mathbf{1}_{n_j}, x\mathbf{w}_j) f_2(v; \boldsymbol{\Theta}) dv \right] \\
= & \sum_{j=1}^{c} tW_j \log \left[\int_{-\infty}^{\infty} \exp(\ell_j^{proper}(\boldsymbol{\theta}; \mathbf{y}_j - v\mathbf{1}_{n_j}, x\mathbf{w}_j) f_2(v; \boldsymbol{\Theta}) dv \right] \\
& + \sum_{j=1}^{c} tW_j \left[C_{1j}(x\mathbf{w}_j; \boldsymbol{\theta}) \right] \\
= & \ell^{proper}(\boldsymbol{\theta}, \boldsymbol{\Theta}; \{\mathbf{y}_j, x\mathbf{w}_j, j = 1, .., c\}; t\mathbf{W}) \\
& + \sum_{j=1}^{c} tW_j \left[C_{1j}(x\mathbf{w}_j; \boldsymbol{\theta}) \right] + C_2(\{x\mathbf{w}_j, j = 1, ..., c\}, t\mathbf{W}; \boldsymbol{\theta}, \boldsymbol{\Theta})
\end{aligned}
$$

Normal distribution - Two stage design

Population model :

- $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma)$
- $Y_{ij} \sim N(\mathbf{x}_{ij}^t \boldsymbol{\beta} - v, \sigma^2), i = 1, ..., n_j$ independent observations with pdf $f_1(y - v; \boldsymbol{\theta})$ given random effect $V_j = v$.
- $\boldsymbol{\Theta} = (\eta)$
- $V_j \sim N(0, \eta^2), j = 1, ..., c$: independent random effects with pdf $f_2(v; \boldsymbol{\Theta}) =$

The model and the within-PSU weighting scheme imply

Sampling distribution :

$(\mathbf{Y}_1, ..., \mathbf{Y}_c)$ are independent vectors with

$$\mathbf{Y}_j \sim N(\mathbf{X}_j^t \boldsymbol{\beta}, \Gamma_j) \qquad \Gamma_j = \frac{\sigma^2}{x} \mathrm{diag}(\mathbf{w}_j)^{-1} + \eta^2 \mathbf{1}\mathbf{1}^T.$$

$$\det(\Gamma_j) = \left(G_j\, \sigma^2/x\right)^{n_j} \frac{n_j \eta^2 + \sigma^2/x}{\sigma^2/x}.$$

$G_j =$ geometric mean of weights $(w_{ij}, i = 1, ..., n_j) = \mathbf{w}_j$.

Normal distribution - Two stage design

Correction term

$$C(\{x\mathbf{w}_j, j=1,...,c\}, t\mathbf{W}; \sigma, \eta) = \sum_j tW_j C_{1j} + C_2 = C_1 + C_2$$

$$\begin{aligned}
2C_1 &= -\sum_j tW_j n_j \left\{(x-1)\log\left(2\pi\sigma^2\right) + [\log(x) + \log(G_j)]\right\} \\
2C_2 &= \sum_j n_j \log(W_j) + (tW_j - 1)\log[\det(\Gamma_j)] \\
&= \sum_j n_j[\log(tW_j) + (tW_j - 1)\log(G_i)] \\
&- (\sum_j n_j)(t-1)\log(\sigma^2/x)] + \sum_j (tW_j - 1)\log\left[\frac{n_j\eta^2 + \sigma^2/x}{\sigma^2/x}\right].
\end{aligned}$$

Normal distribution - Two stage design

- $x = 1$ makes C_1 independent of σ i.e.

  $x = 1 \Longrightarrow \ell_j^{pseudo}$ and ℓ_j^{proper} are equivalent for all j.
- if moreover $t = 1$, C_2 is independent of σ and η in two instances :
  1. if $n_j = n$, then $\Gamma_j = \Gamma$ and $\sum_j W_j = c$,
  $$2C_2 = n \sum_{j=1}^{c} [\log(W_j)]$$
  2. if $W_j = 1$,
  $$2C_2 = 0.$$

In all other cases, the overall log-likelihoods ℓ^{pseudo} and ℓ^{proper} will give different estimates.

Multivariate generalized beta distribution (MGB2)
Two stage design

MGB2 distribution (Yang et al., 2010) :
a set of n random variables $\mathbf{Y} = (Y_1, ..., Y_n)$ conditionally independent given a random scale parameter Θ, with pdf

$$\mathbf{Y}|\{\Theta = \theta\} \sim GG(a, (\theta^{-1/a}\mathbf{b}), p)$$

$\Theta \sim invGa(q)$ with pdf

$$g(\theta; q) = \frac{1}{\Gamma(q)}\theta^{-q}e^{-\theta}\frac{1}{\theta}$$

Graf, Marín and Molina (2018) use this setting in the context of small area estimation.

- Θ :latent area effect
- $\log(\mathbf{b}) = \mathbf{X}\beta$: model on scale
- a, p and q : shape parameters

MGB2 - two stage

Aim : incorporate weights.
Same setting as in the normal case.

- PSU j : sample size $n_j, j = 1, ..., c$,
  canonical weights $\mathbf{w}_j = (w_{ij}, i = 1, ..., n_j)$
  $\log(b_{ij}) = \mathbf{x}_{ij}^t \boldsymbol{\beta}$
- PSU canonical weights : W_j
- x and t scaling factors.
- ℓ_j^{proper} : sum of log-densities $GG(a, (\theta x w_{ij})^{-1/a} b_{ij}, p x w_{ij})$
- $\Theta_j \sim invG(t W_j q)$
- PSU are independent.

MGB2 - two stage

Correction terms

$$
\begin{aligned}
C_{1j}(x) &= n_j(x-1)\log(a) - n_j x \log(\Gamma(p)) + \sum_{i=1}^{n_j} \log(\Gamma(x w_{ij} p)) \\
C_2(t,x) &= c(t-1)\log(a) + \\
&\quad \sum_{j=1}^{c} tW_j \log\left[\frac{\Gamma(x n_j p + q)}{\Gamma(q)\prod_{i=1}^{n_j}[\Gamma(x w_{ij} p)]}\right] - \\
&\quad \sum_{j=1}^{c} \log\left[\frac{\Gamma(tW_j x n_j p + tW_j q)}{\Gamma(tW_j q)\prod_{i=1}^{n_j}[\Gamma(tW_j x w_{ij} p)]}\right] \\
C(t,x) &= \sum_{j=1}^{c} tW_j C_{1j}(x) + C_2(t,x).
\end{aligned}
$$

MGB2 - two stage

- $C(t, x)$ does not depend on a if and only if $t = 1$ and $x = 1$.
- $C(1, 1)$ does not depend on q, if $W_j = 1$.
- $C(1, 1)$ still depends on p and q, if $n_j = n$.
- $C(1, 1)$ still depends on p and q, if $w_{ij} = 1$ but $W_j \neq 1$.

The estimates based on ℓ^{proper} or ℓ^{pseudo} won't coincide, except if all the canonical weights are 1.

Discussion

- Design properties of canonical weights.
- Underestimation of between-cluster variance in Gaussian model mentioned by e.g. Rabe-Hesketh and Skrondal (2006) when the expectation of weighted estimates is computed from the population model.
  It does not occur if the sampling distribution is used.
- Advantage of having a sampling density over a method of moments.
- Simpler than the sampling density based on modeling the weights.