Deriving Bayesian and frequentist estimators from time-invariance estimating equations: a unifying approach

Antonietta Mira
Università della Svizzera italiana, Lugano and University of Insubria, Como

Adrian Baddeley
Curtin University, Perth

BFF4, Harvard University, May 2, 2017
INDEX

• Estimating equations
• Time-invariance estimating equations
• Examples of frequentist time-invariance estimators
• Examples of Bayesian time-invariance estimators
• Godambe-Heyde optimality criterion
• Optimality of MLE and Bayesian estimators
• Discussion
Consider a system of random variables \( X = (X_i : i = 1, \cdots, M) \) with a complex joint distribution, \( f(x|\theta) \), governed by parameter \( \theta \).

Our aim is to estimate \( \theta \) given data \( x \) and possibly prior \( \nu(\theta) \).

**Maximum likelihood**

\[ \hat{\theta} = \arg\max_{\theta} f(x, \theta) \]

**Method of moments**

solve for \( \theta \) in \( \mathbb{E}_{X|\theta} S(X) = S(x) \)

**Bayesian estimator** with absolute error loss function

\[ \hat{\theta} = \arg\max_{\theta} \pi(\theta|x) \]

**Bayesian estimator** with squared error loss function

\[ \hat{\theta} = \mathbb{E}_{\theta|x}[\theta] \]

**Estimating equations**

solve for \( \theta \) in \( g(x, \theta) = 0 \)
UNBIASED ESTIMATING EQUATIONS

We call

\[ g(x, \theta) = 0 \]

an **unbiased frequentist** estimating equation if

\[ \mathbb{E}_{X|\theta}[g(X, \theta)] = 0, \quad \forall \theta \]

the estimator is not necessarily unbiased but is usually consistent.

Similarly, define an **unbiased Bayesian** estimating equation if

\[ \mathbb{E}_{\theta|x}[g(x, \theta)] = 0, \quad \forall x \]

Finally, define an **unbiased general** estimating equation if

\[ \mathbb{E}_{\theta,x}[g(x, \theta)] = 0 \]
Example:

In the **method of moments** we solve

\[ \mathbb{E}_{X|\theta} S(X) = S(x) \]

This is equivalent to solve the E.E.: \( g(x, \theta) = 0 \) where

\[ g(x, \theta) = \mathbb{E}_{X|\theta} S(X) - S(x) \]

This is an **unbiased frequentist E.E.**:

\[
\begin{align*}
\mathbb{E}_{X|\theta}\{g(X, \theta)\} &= \mathbb{E}_{X|\theta}\{\mathbb{E}_{X|\theta}[S(X)] - S(X)\} \\
&= \mathbb{E}_{X|\theta}[S(X)] - \mathbb{E}_{X|\theta}[S(X)] \\
&= 0
\end{align*}
\]
Example:

The **Bayesian estimator** for \( h(\theta) \):

\[
\hat{h}(\theta) = \mathbb{E}_{\theta|X} h(\theta)
\]

can be derived from an E.E.: \( g(x, \theta) = 0 \) where

\[
g(x, \theta) = \mathbb{E}_{\theta|x} h(\theta) - h(\theta)
\]

This is an **unbiased Bayesian E.E.:**

\[
\mathbb{E}_{\theta|x}\{g(x, \theta)\} = \mathbb{E}_{\theta|x}\{\mathbb{E}_{\theta|x}[h(\theta)] - h(\theta)\}
\]
\[
= \mathbb{E}_{\theta|x}[h(\theta)] - \mathbb{E}_{\theta|x}[h(\theta)]
\]
\[
= 0
\]
Example:

**MLE for Exponential Family**
Assume the probability density $f$ has the form

$$f(x|\theta) = Z(\theta) \exp\{\theta^T V(x)\}, \quad x \in \mathcal{X}$$

$\theta \in \Theta \subseteq \mathbb{R}^p$ is the canonical parameter
$V : \mathcal{X} \rightarrow \mathbb{R}^p$ is the canonical sufficient statistics

Given data $x$, under regularity conditions, MLE is the solution of

$$U(\hat{\theta}, x) = 0$$

where $U(\theta, x) = \frac{\partial}{\partial \theta} \log f(\theta, x)$ is the *score function*
For a regular exponential family

\[ U(\theta, x) = V(x) - \mathbb{E}_{X|\theta}[V(X)] \]

**MLE** is the solution of

\[ \mathbb{E}_{X|\theta}[V(X)] = V(x) \]

i.e. method of moments applied to \( V \) and the score function is an unbiased frequentist estimating function:

\[ \mathbb{E}_{X|\theta}[U(\theta, X)] = 0 \]
CHALLENGE

The normalizing constant $Z(\theta)$ may be an intractable function of $\theta$ so that the likelihood cannot be maximized explicitly:

$$\frac{1}{Z(\theta)} = \int \exp\{\theta^T V(x)\} \, dx$$

$$\mathbb{E}_{X|\theta}[V(X)] = \frac{\partial}{\partial \theta} \log Z(\theta)$$

and the score function cannot be evaluated.

Possible remedies
- approximate $Z(\theta)$ analytically
- approximate $Z(\theta)$ by simulation
- replace score by another estimating function
Example:

Pseudo-Likelihood Estimator

Suppose $X = (X_1, \cdots, X_M)$ is a vector of discrete random variables. Define

$$PL(\theta, x) = \prod_{i=1}^{M} P_{\theta}\{X_i = x_i | X_{-i} = x_{-i}\}$$

Pseudo-Likelihood estimator:

$$\tilde{\theta} = \arg\max_{\theta \in \Theta} PL(\theta, x)$$

Rationale:

- PL HD good approximation to the LHD if $X_i$ approx. independent
- PL HD usually very simple and tractable
- MPLE is faster to compute than MLE by a factor of $10^2 - 10^4$
- MPLE is consistent and asymptotically normal
- MPLE can be biased and inefficient in small samples
Example:

Pseudo-Likelihood Estimator for Exponential Family

Suppose $X = (X_1, \cdots, X_M)$ has regular exponential family density. Then the MPLE normal equations are:

$$
\frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_{X|\theta}[V(X)|X_{-i} = x_{-i}] = V(x)
$$

This is a frequentist unbiased EE.
Example:
Bayesian Estimator for Exponential Family

If the posterior belongs to the exponential family

$$\pi(\theta | t(x)) = Z(t)^{-1} \exp \{ t^T \phi(\theta) \}$$

$$\phi = \text{sufficient statistic}$$
$$t = t(x) = \text{canonical parameter}$$
$$W(\theta, t) = \text{derivative of log-posterior} = \frac{\partial}{\partial t} \log \pi(\theta, t) = \text{B-Score}$$

For a regular exponential family:

$$W(\theta, t) = \phi(\theta) - \mathbb{E}_{\theta | x} [\phi(\theta)]$$

**BAYES ESTIMATOR** of $\phi(\theta)$: solution of $W(\theta, t) = 0$

$W(\theta, t)$ is an unbiased Bayesian estimating function:

$$\mathbb{E}_{\theta | x} [W(\theta, t(x))] = 0$$
Time-invariance Estimating Equations

(frequentist)

A natural way to understand and simulate a highly structured random process \( X \), with sample space \( \mathcal{X} \), is as the equilibrium distribution of a Markov process \( Y = (Y_n, n = 1, 2 \ldots) \) with states in \( \mathcal{X} \).

<table>
<thead>
<tr>
<th>( X )</th>
<th>( Y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>random pack of cards</td>
<td>shuffling process</td>
</tr>
<tr>
<td>Poisson distribution</td>
<td>birth-and-death process</td>
</tr>
<tr>
<td>Markov random field</td>
<td>Gibbs sampler</td>
</tr>
<tr>
<td>Gibbs point process</td>
<td>spatial birth-death process</td>
</tr>
</tbody>
</table>

- \( Y_n \) is a mathematical fiction in the original context
- The “time” variable, \( n \), is not part of the original system \( X \)
- There are many alternative choices of \( Y \) for each \( X \) (MCMC)
INFINITESIMAL GENERATOR of Markov Chains:

Given a Markov chain $Y_n$ with values on $\mathcal{X}$, the generator $A$ of $Y_n$ is

$$(AS)(x) = \mathbb{E}[S(Y_{n+1}) - S(Y_n) \mid Y_n = x]$$

$$= \mathbb{E}[S(Y_{n+1}) \mid Y_n = x] - S(x)$$

for any $S : \mathcal{X} \rightarrow \mathbb{R}^d$

If $\mathcal{X}$ is finite and $Y$ has transition probabilities $P(x, y)$:

$$(AS)(x) = \sum_{y \in \mathcal{X}} [S(y) - S(x)]P(x, y)$$

$$= \sum_{y \in \mathcal{X}} [S(y)]P(x, y) - S(x)$$

So $A = P - I$
STANDARD RESULT:

If $X \sim \pi$ (the stationary distribution of the MC) then

$$E_\pi[(AS)(X)] = E_\pi[E[S(Y_{n+1}) - S(Y_n)|Y_n = X]]$$
$$= E_\pi[E[S(Y_{n+1})|Y_n = X]] - E_\pi[S(X)]$$
$$= E_\pi[S(Y_{n+1})] - E_\pi[S(Y_n)]$$
$$= 0$$

for essentially all $S$

If $\mathcal{X}$ is finite and $\pi$ is stationary distribution for $Y_n$ i.e. $\pi P = \pi$ then

$$\pi A = \pi (P - I) = \pi P - \pi = 0$$
NEW (unifying) ESTIMATION FRAMEWORK

For each $\theta$, represent the distribution of $X$ under $\theta$, $f(x|\theta)$, as the equilibrium distribution of some $Y^{(\theta)} = (Y_n^{(\theta)})$

Let $A_\theta = \text{infinitesimal generator of } Y^{(\theta)}$

Choose a statistic $S = S(X)$

Given observed data $X = x$, estimate $\theta$ as the solution (if $\exists$ !) $\hat{\theta}_T$ of

$$(A_\theta S)(x) = 0$$

This is an unbiased frequentist estimating equation for $\theta$

$\hat{\theta}_T = \text{TIME-INVARIANCE ESTIMATOR}$

Degrees of freedom: $Y$ and $S$
<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
<th>$S(x)$</th>
<th>$\hat{\theta}_T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>anything</td>
<td>i.i.d.</td>
<td>any</td>
<td>M-of-M</td>
</tr>
<tr>
<td>exp. family</td>
<td>i.i.d.</td>
<td>$V(x)$</td>
<td>MLE</td>
</tr>
<tr>
<td>exp. family</td>
<td>Gibbs sampler</td>
<td>$V(x)$</td>
<td>MPLIE</td>
</tr>
<tr>
<td>exp. family</td>
<td>Langevin Diff</td>
<td>$V(x)$</td>
<td>MLE</td>
</tr>
<tr>
<td>Poisson dist</td>
<td>Birth-Death</td>
<td>$x$</td>
<td>MLE</td>
</tr>
<tr>
<td>continuous r.v.</td>
<td>OU diffusion</td>
<td>any</td>
<td>Variational Estimator</td>
</tr>
<tr>
<td>$X$</td>
<td>$Y$</td>
<td>$S(x)$</td>
<td>$\hat{\theta}_T$</td>
</tr>
<tr>
<td>-------------------</td>
<td>----------------------</td>
<td>---------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>anything</td>
<td>i.i.d.</td>
<td>any</td>
<td>M-of-M</td>
</tr>
<tr>
<td>exp. family</td>
<td>i.i.d.</td>
<td>$V(x)$</td>
<td>MLE</td>
</tr>
<tr>
<td>exp. family</td>
<td>Gibbs sampler</td>
<td>$V(x)$</td>
<td>MPLE</td>
</tr>
<tr>
<td>exp. family</td>
<td>Langevin Diff</td>
<td>$V(x)$</td>
<td>MLE</td>
</tr>
<tr>
<td>Poisson dist</td>
<td>Birth-Death</td>
<td>$x$</td>
<td>MLE</td>
</tr>
<tr>
<td>continuous r.v.</td>
<td>OU diffusion</td>
<td>any</td>
<td>Variational Estimator</td>
</tr>
</tbody>
</table>
Example: Method of moments and MLE

Let $X$ be a random element of any space $\mathcal{X}$
Let $Y_n$ be i.i.d. copies of $X_n$
Then $Y$ has infinitesimal generator

$$(AS)(x) = \mathbb{E}[S(X_{n+1}) - S(X_n) \mid X_n = x] = \mathbb{E}[S(X)] - S(x)$$

and $\hat{\theta}_T$ is the solution of

$$\mathbb{E}_\theta [S(X)] = S(x)$$

i.e. this leads to the method-of-moments estimator

If $X$ has a distribution from an exponential family
take $S =$ canonical sufficient statistics
and $Y =$ iid sampling from $f(x \mid \theta)$
then time-invariance EE gives the normal equations for the MLE
<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>$S(x)$</th>
<th>$\hat{\theta}_T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>anything</td>
<td>i.i.d.</td>
<td>any</td>
<td>M-of-M</td>
</tr>
<tr>
<td>exp. family</td>
<td>i.i.d.</td>
<td>$V(x)$</td>
<td>MLE</td>
</tr>
<tr>
<td><strong>exp. Family</strong></td>
<td>Gibbs sampler</td>
<td>$V(x)$</td>
<td><strong>MPLE</strong></td>
</tr>
<tr>
<td>exp. family</td>
<td>Langevin Diff</td>
<td>$V(x)$</td>
<td>MLE</td>
</tr>
<tr>
<td>Poisson dist</td>
<td>Birth-Death</td>
<td>$x$</td>
<td>MLE</td>
</tr>
<tr>
<td>continuous r.v.</td>
<td>OU diffusion</td>
<td>any</td>
<td>Variational Estimator</td>
</tr>
</tbody>
</table>
Example: Maximum Pseudo-LHD for Exp. Family

Given a finite system of r.v. \( X = (X_1, \cdots, X_M) \)
Take \( Y \) to be single site random scan Gibbs sampler
The generator of \( Y \) is:

\[
(AS)(x) = \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_{X|\theta}[S(X)|X_{-i} = x_{-i}] - S(x)
\]

If we choose \( S = V \) and
\( f(x|\theta) \) belongs to the exponential family,
then the time-invariance EE is equivalent to the maximum pseudo-likelihood normal equations
<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
<th>$S(x)$</th>
<th>$\hat{\theta}_T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>anything</td>
<td>i.i.d.</td>
<td>any</td>
<td>M-of-M</td>
</tr>
<tr>
<td>exp. family</td>
<td>i.i.d.</td>
<td>$V(x)$</td>
<td>MLE</td>
</tr>
<tr>
<td>exp. family</td>
<td>Gibbs sampler</td>
<td>$V(x)$</td>
<td>MPLE</td>
</tr>
<tr>
<td>exp. family</td>
<td>Langevin Diff</td>
<td>$V(x)$</td>
<td>MLE</td>
</tr>
<tr>
<td><strong>Poisson dist</strong></td>
<td>Birth-Death</td>
<td>$x$</td>
<td>MLE</td>
</tr>
<tr>
<td>continuous r.v.</td>
<td>OU diffusion</td>
<td>any</td>
<td>Variational</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Estimator</td>
</tr>
</tbody>
</table>
Example: MLE and MoM for Poisson distribution

Let $X \sim \text{Poisson}(\lambda)$, $\lambda > 0$, unknown
Thus $\mathcal{X} = \mathbb{N} = \{0, 1, 2, \ldots\}$ and $\Theta = \mathbb{R}_+$
We observe a single realisation $x$ of $X$
Let $Y_n$ be the birth-death process with transition rates

$$r(x, x + 1) = \lambda \quad \text{and} \quad r(x, x - 1) = x$$

Then for any $S : \mathbb{N} \rightarrow \mathbb{N}$

$$(\mathcal{A}_\lambda S)(x) = \lambda [S(x + 1) - S(x)] + x [S(x - 1) - S(x)]$$

Take $S(x) \equiv x$; then $$(\mathcal{A}_\lambda S)(x) = \lambda - x$$
and the time-invariance estimator is

$$\hat{\lambda}_T = x$$

which is also the MLE and method-of-moments estimator
<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
<th>$S(x)$</th>
<th>$\hat{\theta}_T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>M. random field</td>
<td>Gibbs sampler</td>
<td>$V(x)$</td>
<td>MPLE</td>
</tr>
<tr>
<td>M. random field</td>
<td>block-update Gibbs</td>
<td>$V(x)$</td>
<td>Generalized MPLE</td>
</tr>
<tr>
<td>point process</td>
<td>spatial birth-death</td>
<td>$V(x)$</td>
<td>MPLE</td>
</tr>
<tr>
<td>point process</td>
<td>multiple birth-death</td>
<td>$V(x)$</td>
<td>Generalized MPLE</td>
</tr>
<tr>
<td>point process</td>
<td>spatial birth-death</td>
<td>any</td>
<td>Takacs-Fiksel</td>
</tr>
<tr>
<td>random censoring</td>
<td>random lifetime</td>
<td>e.d.f.</td>
<td>Reduced sample estimator</td>
</tr>
</tbody>
</table>
**TYPE I time-invariance EE:**

\[(\mathcal{A}_\theta S)(x) = \mathbb{E}[S(X_{n+1}) - S(X_n) \mid X_n = x]\]

where \(S: \mathcal{X} \rightarrow \mathbb{R}^d\)

**TYPE II time-invariance EE:**

From the connection with Stein-Chen method of distributional approximation we can define

\[(\mathcal{F} T)(x) = \mathbb{E}[T(X_n, X_{n+1}) \mid X_n = x]\] (1)

where \(T: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}\) is **antisymmetric**: \(T(x, y) = -T(y, x)\)

and the Markov chain is in discrete time and reversible wrt \(\pi\).

Then we have

\[\mathbb{E}_\pi (\mathcal{F} T)(X) = 0\]

so this is another source of frequentist unbiased EE

Type II is a generalization of Type I: take \(T(x, y) = S(y) - S(x)\)
Missing observations

observable data = \( X \)
unobservable data = \( Z \)
full data = \((X, Z)|\theta \sim f(X, Z|\theta)\)

the time-invariance type I and II EE becomes:

\[
(A_\theta S)(x, z) = 0
\]
\[
(F_\theta T)(x, z) = 0
\]

\( x = \text{observed} \) realization of \( X \)
\( z = \text{unobservable} \) realization of \( Z \)

These equations give rise to usable estimators only if \textit{the left hand sides do not depend on } z
This can be obtained taking $Y$ to be a Gibbs sampler:

$$
(x, z) \xrightarrow{\mathbb{P}(z'|x)} (x, z') \xrightarrow{\mathbb{P}(x'|z')} (x', z')
$$

$$(A_\theta S)(x, z) = 0$$

$$
\int_Z \int_X [S(x', z') - S(x, z)] \mathbb{P}_\theta(dx'|z') \mathbb{P}_\theta(dz'|x) = 0
$$

$$
\mathbb{E}_\theta[U_\theta(Z)|x] - S(x, z) = 0
$$

where $U_\theta(z) = \mathbb{E}_\theta[S(X, z)|z]$. This is an unbiased EE wrt the joint: $(X, Z)|\theta$

If $S(x, z) = V(x)$ in an exp. fam. then the EE becomes:

$$
\mathbb{E}_\theta[U_\theta(Z)|x] = V(x)
$$

and it is an unbiased EE wrt the marginal: $X|\theta$
We may also want to predict $Z$ from $X$

Then take $S(x, z) = V(z)$

It follows that:

$$U_\theta(z) = \mathbb{E}_\theta[S(X, z) \mid z] = V(x)$$

The EE becomes:

$$\mathbb{E}_\theta[V(Z) \mid x] = V(z)$$

and is an unbiased EE wrt the conditional: $Z \mid x, \theta$
Bayesian Time-Invariance E.E. on $\Theta$:

Take the process $Y$ have **state space** $\Theta$, instead of $\mathcal{X}$
and be **stationary wrt** $\pi(\theta|x)$, instead of $f(x|\theta)$

**TYPE I EE:**
Let $Y$ be **i.i.d. sampling** from the posterior
Take:

$$S(\theta) = g(\theta)$$ the function we want to estimate

Then the time-invariance E.E. gives Bayesian estimator for $g(\theta)$

**TYPE II EE:**
Let $Y$ be an **independence MH** with **proposal** = prior
Take:

$$T(\theta, \theta') = \frac{g(\theta')\pi(\theta|x)}{\nu(\theta')\alpha(\theta', \theta)} - \frac{g(\theta)\pi(\theta'|x)}{\nu(\theta)\alpha(\theta, \theta')}$$

Then the time-invariance E.E. gives Bayesian estimator for $g(\theta)$
General Time-Invariance Estimating Equations:

Suppose $X$ and $\theta$ are both random:

\[
\begin{align*}
\theta & \sim \nu(\theta) \\
X \mid \theta & \sim f(x \mid \theta) \\
\theta \mid x & \sim \pi(\theta \mid x)
\end{align*}
\]

Construct a process on $(X, \Theta)$ stationary wrt $\nu(\theta)f(x \mid \theta)$.

For example consider the Gibbs sampler:

\[
(x, \theta) \xrightarrow{\pi(\theta'|x)} (x, \theta') \xrightarrow{f(x'|\theta')} (x', \theta')
\]

Then **Type I EE**

\[
(AS)(x, \theta) = \int_{\Theta} \int_{X} S(x', \theta') f(x' \mid \theta') \pi(\theta' \mid x) \, dx' \, d\theta' - S(x, \theta)
\]

which is an unbiased EE wrt the joint of $(X, \Theta) : \nu(\theta)f(x \mid \theta)$.
Unifying framework

$Y$ process on the $\mathcal{X}$ space $\rightarrow$ TI frequentist EE
$Y$ process on the $\Theta$ space $\rightarrow$ TI Bayesian EE
$Y$ process on the $(\mathcal{X}, \Theta)$ space $\rightarrow$ TI general EE

$Y$ could be used to study properties of the resulting estimators

Eaton’s $\Theta$-chain + Hobert & Robert’s $\mathcal{X}$-chain: studied to
determine whether the formal Bayes estimate is $\nu$-admissible
If either of these chains is null recurrent the Bayes estimate is
$\nu$-admissible
The $\mathcal{X}$, the $\Theta$ and the $(\mathcal{X}, \Theta)$-chains are always either
all (positive) recurrent or all transient
$m(x)$ is proper $\iff$ prior is proper
ADMISSIBILITY $\iff$ RECURRENCE

Brown (1971) showed that the (least squares) best invariant estimator of the mean of a multivariate normal distribution is admissible if and only if an associated diffusion is recurrent.

Table: Updating regimes on $\Theta$ and resulting estimators

<table>
<thead>
<tr>
<th>Stat. dist.</th>
<th>Markov process</th>
<th>Function</th>
<th>Type</th>
<th>Estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi(\theta</td>
<td>x)$</td>
<td>i.i.d. from $\pi(\theta</td>
<td>x)$</td>
<td>$S(\theta) = g(\theta)$</td>
</tr>
<tr>
<td>$\pi(\theta</td>
<td>x)$</td>
<td>M-H with prop. $\nu(\theta)$</td>
<td>$T(\theta, \theta') = (2)$</td>
<td>II</td>
</tr>
<tr>
<td>$\nu(\theta)$ Eaton’s chain</td>
<td>$f(x'</td>
<td>\theta)\pi(\theta'</td>
<td>x')$</td>
<td>$T(\theta, x) = h(x) - g(\theta)$</td>
</tr>
</tbody>
</table>

$$T(\theta, \theta') = \frac{g(\theta')\pi(\theta|x)}{\nu(\theta)\alpha(\theta', \theta)} - \frac{g(\theta)\pi(\theta'|x)}{\nu(\theta')\alpha(\theta, \theta')}$$ (2)

These EE are unbiased wrt the marginal of $\theta$
Table: Updating regimes on $\mathbf{x}$ and resulting estimators

<table>
<thead>
<tr>
<th>Stat. dist.</th>
<th>Markov process</th>
<th>Function</th>
<th>Type</th>
<th>Estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f(x</td>
<td>\theta)$</td>
<td>i.i.d. from $f(x</td>
<td>\theta)$</td>
<td>$S(x) = h(x)$</td>
</tr>
<tr>
<td>$f(x</td>
<td>\theta)$</td>
<td>M-H with prop. $m(x)$</td>
<td>$T(x, x') = (3)$</td>
<td>II</td>
</tr>
<tr>
<td>$m(x)$</td>
<td>Hobert’s chain $\pi(\theta'</td>
<td>x)f(x'</td>
<td>\theta')$</td>
<td>$T(\theta, x) = g(\theta) - h(x)$</td>
</tr>
</tbody>
</table>

$$T(x, x') = \frac{h(x')f(x|\theta)}{m(x)\alpha(x', x)} - \frac{h(x)f(x'|\theta)}{m(x')\alpha(x, x')} \quad (3)$$

These EE are unbiased wrt the marginal of $X$.
Table: Updating regimes on \((X, \Theta)\) and resulting estimators

<table>
<thead>
<tr>
<th>Stat. dist.</th>
<th>Markov process</th>
<th>Function</th>
<th>Estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>(f(x</td>
<td>\theta)\nu(\theta))</td>
<td>(\pi(\theta'</td>
<td>x)f(x'</td>
</tr>
<tr>
<td>(\pi(\theta</td>
<td>x)m(x))</td>
<td>(f(x'</td>
<td>\theta)\pi(\theta'</td>
</tr>
</tbody>
</table>

These EE are unbiased wrt the joint \((X, \theta)\)
Table: Updating regimes and resulting estimators on $(\mathcal{X}, \Theta)$

<table>
<thead>
<tr>
<th>Stat. dist.</th>
<th>Markov process</th>
<th>Function</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi(\theta</td>
<td>x)m(x)$</td>
<td>$x'_i \sim f(x'_i</td>
<td>x_{-i}, \theta)$, $\theta' \sim \pi(\theta'</td>
</tr>
<tr>
<td>$f(x</td>
<td>\theta)\nu(\theta)$</td>
<td>$\theta'_j \sim \pi(\theta'_j</td>
<td>\theta_{-j}, x)$, $x' \sim f(x'</td>
</tr>
<tr>
<td>$\pi(\theta</td>
<td>x)m(x)$</td>
<td>$x' \sim \text{block Gibbs sampler}$, $\theta' \sim \pi(\theta'</td>
<td>x')$</td>
</tr>
<tr>
<td>$f(x</td>
<td>\theta)\nu(\theta)$</td>
<td>$\theta \sim \text{block Gibbs sampler}$, $x \sim f(x'</td>
<td>heta')$</td>
</tr>
</tbody>
</table>
This approach yields many different estimators (old and new) depending on the choices of $Y$ and $S$

**WHICH CHOICE OF $Y$ AND $S$ IS “BEST”?**
(frequentist) Many of these estimators can be interpreted as maximising a variant of the likelihood.

If the chain $Y$ has good performance as an MCMC sampler delivers efficient MCMC estimators explores the state space rapidly

then $(AS)(x)$ should be close to efficient score $U(x, \theta) = \mathbb{E}_\theta S(X) - S(x)$ so that the time-invariance estimator should be close to the MLE.
Godambe-Heyde asymptotic variance optimality criterion

First order approximation to the asymptotic variance of the TI estimator for $\theta$, obtained using $A_\theta$ and the statistic $V$:

$$GH(\theta) = \frac{\text{Var}_\theta [(A_\theta V)(X)]}{(\mathbb{E}_\theta \frac{\partial}{\partial \theta} (A_\theta V)(X))^2}$$

in general

$$= \frac{\mathbb{E}_\theta [(A_\theta V)(X)^2]}{\mathbb{E}_\theta [V(X)(A_\theta V)(X)]^2}$$

for exp. fam.
GODAMBE-HEYDE ORDERING

Take $f(x|\theta) \in$ exponential family with parameter space $\Theta$
Let $S = V =$ canonical sufficient statistics
Let $Y$ and $Z$ be two stochastic processes stationary wrt $f(x|\theta)$
We say that $\hat{\theta}_Y \succeq_{GH} \hat{\theta}_Z$ if

$$GH(\theta)_Y \leq GH(\theta)_Z, \quad \forall \theta \in \Theta$$

By Cauchy-Schwarz:

$$GH(\theta) \geq \frac{1}{\text{Var}_\theta[V(X)]}$$
MLE achieves the bound since, for iid sampling:

$$\mathbb{E}[(AV)(X)^2] = \text{Var}[V(X)] = \mathbb{E}[V(X) (AV)(X)]$$

and hence

$$GH(\theta) = \frac{1}{\text{Var}[V(X)]}$$

Thus \textbf{MLE = GH-OPTIMAL} amongst all time-invariance estimators for exponential families derived from $V =$ sufficient statistics
Bayesian estimators achieves the GH lower bound since can be obtained by taking $Y$ to be iid sampling from the posterior

Thus BAYES estimator $= \text{GH-OPTIMAL}$ amongst all time-invariance estimators for exponential families derived from $V = \text{sufficient statistics}$
ORDERINGS for MCMC

EFFICIENCY ORDERING

\( \mathbb{P} \) is **uniformly** more efficient than \( \mathbb{Q} \) for MCMC purposes if

\[
\mathbb{P} \succeq_{E} \mathbb{Q} \iff \text{Var}(S, \mathbb{P}) \leq \text{Var}(S, \mathbb{Q}) \quad \forall S \in L^2(\pi)
\]

where \( \text{Var}(S, \mathbb{P}) = \text{asympt. variance of MCMC estimator for } E_\pi S \)

COVARIANCE ORDERING  Mira and Geyer (2000)

\( \mathbb{P} \) better than \( \mathbb{Q} \) in covariance ordering if:

\[
\mathbb{P} \succeq_{C} \mathbb{Q} \iff \text{Cov}_\pi [S(X_0), S(X_1)] \leq \text{Cov}_\pi [S(Y_0), S(Y_1)] \quad \forall S \in L^2(\pi)
\]

THEOREM: \( \mathbb{P} \succeq_{E} \mathbb{Q} \iff \mathbb{P} \succeq_{C} \mathbb{Q} \)
\( f(x|\theta) = \) exponential family with sufficient statistic \( V \)

Let \( Y_\theta \) and \( Z_\theta \) be MC reversible wrt \( f(x|\theta) \)

Let \( \hat{\theta}_Y \) and \( \hat{\theta}_Z \) be corresponding TI estimators for the statistic \( V \)

If

1. \( Y_\theta \) dominates \( Z_\theta \) in the covariance ordering, for each \( \theta \); and

2. \( \text{Var}_\theta[(A_{Y_\theta} V)(X)] \leq \text{Var}_\theta[(A_{Z_\theta} V)(X)] \),

then

\[ \hat{\theta}_Y \preceq_{GH} \hat{\theta}_Z \]

i.e. \( Y \) is better if it has both

- stronger negative correlation \( \mathbb{E}[V(X)(AV)(X)] \)
- weaker positive correlation \( \mathbb{E}[V(X)(A^2 V)(X)] \)

than the corresponding terms for \( Z \)
EXAMPLE: Geometric distribution

\[ \pi(x) = (1 - p)p^{x-1} \text{ for } x = 1, 2, \cdots \text{ and } 0 \leq p \leq 1 \]

Take \( S(x) = V(x) = x \), the canonical sufficient statistic

Compare:

- **iid sampler**: \( \hat{\rho}_{\text{iid}} = 1 - \frac{1}{x} = \text{MLE} = \text{MM} \)

- **truncated iid sampler**: if \( X_t = x \), a jump to \( y \) is proposed with probability \( \pi(y) \); the proposal is accepted unless \( |x - y| > c \)
  
  For \( c = 1 \): \( \hat{\rho}_1 = 0 \) if \( x > 1 \) and \( \hat{\rho} = 0, 1 \) if \( x = 1 \)

- **MH sampler** with reflecting symmetric RW proposal:
  
  \( \hat{\rho}_{\text{MH}} = 1 \{ x > 1 \} \)

\[
\begin{align*}
\text{iid} & \succeq_P \text{tiid}_c & \forall c \\
\text{tiid}_c & \succeq_P \text{tiid}_{c'} & \forall c' < c \\
\text{MH} & \succeq_P \text{tiid}_1 & p \geq 0.5
\end{align*}
\]
Values of Godambe-Heyde ratio plotted against $p$
Denominators of the Godambe-Heyde ratio
Numerator of the Godambe-Heyde criterion

![Graph showing the numerators of the Godambe-Heyde criterion for different distributions: iid, tild, and MH. The x-axis represents the parameter p ranging from 0.0 to 1.0, and the y-axis represents the numerator values on a logarithmic scale from $10^{-5}$ to $10^3$. The graph includes three curves, each representing a different distribution type.](image-url)
Godambe-Heyde ratio for the tiid with \( c = 1, 2, 3, 10, \infty \)
GH ratio for tiid with $c = 3$ (solid) and the MH (dotted)
closeness of the proposal \( q(\cdot) \) to the target distribution
\[ \downarrow \]
closeness of the estimating fct \( A_\theta V(x) \) to the efficient score \( U(x, \theta) \)
\[ \downarrow \]
high GH-efficiency of the time-invariance estimator
MPLE is “linearisation” of MLE

A standard result:

$$\exp\{tA\} = P_t$$

i.e.

$$(\exp\{tA\} h)(x) = \mathbb{E}_\theta[h(Y_t) \mid Y_0 = x]$$

If $Y$ converges in distribution to $X$ then we have

$$\lim_{t \to \infty} \exp\{tA\} V(x) = \mathbb{E}_\theta V(X)$$

thus the MLE is the solution of the normal equations

$$\lim_{t \to \infty} \exp\{tA_\theta\} V(x) = V(x)$$

or

$$\lim_{t \to \infty} (\exp\{tA_\theta\} - I) V(x) = 0$$

The pseudo-likelihood estimating equations are

$$A_\theta V(x) = 0$$

and we can think of $A_\theta$ as a first order approximation to the series

$$\exp\{tA\} - I$$
TI-EE are particularly relevant when:

- model of interest is highly structured
- likelihood or the posterior are not known analytically but can be expressed as equilibrium distributions of an associated Markov processes

This is the same setting when MCMC and MPLE are helpful

We can think of TI-EE as the “middle way”

**Statistical performance**: invariance, consistency, asymptotic normality and optimality discussed in Baddeley (2000)

**Spatial processes**: Kurtz and Li (2004) give LLN + CLT for TI-Estimators
Conclusions

- TI-EE involve an arbitrary choice of $Y$ and $S$
  - $Y$ is a mathematical fiction
  - natural choices of $S$ and $Y$ often yield known estimators (BE, MLE, M-of-M, MPLE, Takacs-Fiksel, variational est...)
- TI-EE = common structure for different estimation techniques
- TI-EE = general recipe for deriving new estimators
- TI-EE = general framework for comparing estimators
- Computational complexity of $\hat{\theta}$ depends on $Y$ transition structure
- Efficiency of time-invariance estimators is related to the quality of $Y$ as an MCMC sampler
Open questions

- How can we incorporate the loss fct in Bayesian TI-EE?
- Which $Y$ and $S$ give TI-EE that result in Stein estimator?
- Derive Kaplan-Meier estimator as a time-invariance estimator.
- How to study large sample properties of TI-EE?
- Conjecture that large sampler properties of time-invariance estimators mimic those of MLE, MPLE, BE
- How to study large-domain limit theory for spatial models?
- How to handle random effects?