Wasserstein Consensus for Bayesian Sample Size Determination

Tullia Padellini

tulliapadellini.github.io

Milano – 11 October 2019

Dipartimento di Scienze Statistiche
Pre-sperimental problem consisting of choosing the size of the sample, $n$, typically trying to minimize uncertainty under some cost constraint.
Pre-sperimental problem consisting of choosing the size of the sample, \( n \), typically trying to minimize uncertainty under some cost constraint.

In **Clinical Trials** this translates into:

- **Cost** - every patient is "precious" for both ethics and finances
- **Uncertainty** - we cannot risk introducing a dangerous treatment
Pre-sperimental problem consisting of choosing the size of the sample, $n$, typically trying to minimize uncertainty under some cost constraint.

In Clinical Trials this translates into:

- **Cost** - every patient is "precious" for both ethics and finances
- **Uncertainty** - we cannot risk introducing a dangerous treatment

**GOAL:** find a sample size that induces agreement between different parties
The (bayesian) state of the art
the main ingredients

- **Analysis Prior** $\pi_A(\theta)$:
  models pre-experimental information to be used to obtain the **posterior distribution**

- **Design Prior** $\pi_D(\theta)$:
  models uncertainty on the experiment to be used to obtain the **predictive distribution**
The (bayesian) state of the art

the main ingredients

- **Analysis Prior** $\pi_A(\theta)$:
  models pre-experimental information to be used to obtain the
  **posterior distribution**

- **Design Prior** $\pi_D(\theta)$:
  models uncertainty on the experiment to be used to obtain the
  **predictive distribution**

Select $n$ in order to satisfy some inferential goal, to be formalized in terms
of a summary of the posteriors

$$
\rho_{\pi_A}(\theta|y_n) = \int g(\theta)\pi_A(\theta|y_n)d\theta
$$
The design predictive distribution $m_D(y)$ removes the dependency of $\rho_{\pi_A}(\theta | y_n)$ from the observed sample $y_n$. 

**PEC** - Predictive Expectation Criterion

$$e(n) = E_{m_D}[\rho_{\pi_A}(\theta | Y_n)]$$

$$n^\ast = \min\{n \in \mathbb{N}: e(n) > \tau\}$$

**PPC** - Predictive Probability Criterion

$$p(n) = P_{m_D}[\rho_{\pi_A}(\theta | X_n) > \tau]$$

$$n^\ast = \min\{n \in \mathbb{N}: p(n) > \tau\}$$

$\tau$ and $\tau'$ are clinically relevant thresholds and depend on the problem.
The design predictive distribution $m_D(y)$ removes the dependency of $\rho_{\pi_A}(\theta|y_n)$ from the observed sample $y_n$

- **PEC** - Predictive Expectation Criterion

  $$e(n) = \mathbb{E}_{m_D}[\rho_{\pi_A}(\theta|Y_n)]$$

  $$n^* = \min\{n \in \mathbb{N} : e(n) > \eta\}$$
The design predictive distribution $m_D(y)$ removes the dependency of $\rho_{\pi_A}(\theta|y_n)$ from the observed sample $y_n$

> **PEC - Predictive Expectation Criterion**

$$e(n) = \mathbb{E}_{m_D}[\rho_{\pi_A}(\theta|Y_n)]$$

$$n^* = \min\{n \in \mathbb{N} : e(n) > \eta\}$$

> **PPC - Predictive Probability Criterion**

$$p(n) = \mathbb{P}_{m_D}[\rho_{\pi_A}(\theta|X_n) > \gamma]$$

$$n^* = \min\{n \in \mathbb{N} : p(n) > \eta\}$$

$\rho$ and $\eta$ are clinically relevant thresholds and depend on the problem.
The design predictive distribution $m_D(y)$ removes the dependency of $\rho_{\pi_A}(\theta|y_n)$ from the observed sample $y_n$

- **PEC** - Predictive Expectation Criterion

$$e(n) = \mathbb{E}_{m_D}[\rho_{\pi_A}(\theta|Y_n)] \quad n^* = \min\{n \in \mathbb{N} : e(n) > \eta\}$$

- **PPC** - Predictive Probability Criterion

$$p(n) = \mathbb{P}_{m_D}[\rho_{\pi_A}(\theta|X_n) > \gamma] \quad n^* = \min\{n \in \mathbb{N} : p(n) > \eta\}$$

$\eta$ and $\gamma$ are clinically relevant thresholds and depend on the problem.
Multiple priors
when should we look for "consensus"?

- diverging expert opinions
- multiple scenarios to take into account
- data from previous studies
Multiple priors
when should we look for "consensus"?

- diverging expert opinions
- multiple scenarios to take into account
- data from previous studies

**Community of priors problem**: how to combine multiple sources of pre-sperimental information into the analysis?
Aggregate multiple priors into one and then use the approach of your likings.

Will the $i$-th clinician believe us?
Aggregate multiple priors into one and then use the approach of your likings.

\[ \pi_1(\theta), \ldots, \pi_K(\theta) \]
The standard Solution
mixtures of priors

Aggregate multiple priors into one and then use the approach of your likings.

\[ \pi_1(\theta), \ldots, \pi_K(\theta) \xrightarrow{\text{ Aggregate }} \pi_A(\theta) = \sum_{i=1}^{K} \omega_{i} \pi_i(\theta) \]

Will the i-th clinician believe us?

The standard Solution
mixtures of priors

Aggregate multiple priors into one and then use the approach of your likings.

\[ \pi_1(\theta), \ldots, \pi_K(\theta) \quad \xrightarrow{\text{Aggregate}} \quad \pi_A(\theta) = \sum_{i=1}^{K} \omega_{o,i} \pi_i(\theta) \]

\[ \pi_A(\theta|y_n) = \sum_{i=1}^{K} \frac{\omega_{o,i} m_i(y_n)}{\sum_{r=1}^{K} \omega_{o,r} m_r(y_n)} \times \pi_i(\theta|y_n) \]

Will the \( i \)-th clinician believe us?

The standard Solution
mixtures of priors

Aggregate multiple priors into one and then use the approach of your likings.

$$\pi_1(\theta), \ldots, \pi_K(\theta) \quad \rightarrow \quad \pi_A(\theta) = \sum_{i=1}^{K} \omega_{o,i} \pi_i(\theta)$$

$$\pi_A(\theta|y_n) = \sum_{i=1}^{K} \frac{\omega_{o,i} m_i(y_n)}{\sum_{r=1}^{K} \omega_{o,r} m_r(y_n)} \times \pi_i(\theta|y_n)$$

Will the $i$-th clinician believe us?

Two experts “agree” if their inferential conclusions are the same.
Our Solution
enforcing “consensus” between sources

› (possibly) conflicting priors $\pi_1, \pi_2$
› resulting posteriors $\pi_{1,y}, \pi_{2,y}$

Two experts “agree” if their inferential conclusions are the same, hence if their **posterior distributions are close enough**.
Our Solution
enforcing “consensus” between sources

- (possibly) conflicting priors \(\pi_1, \pi_2\)
- resulting posteriors \(\pi_{1,y}, \pi_{2,y}\)

Two experts “agree” if their inferential conclusions are the same, hence if their **posterior distributions are close enough**.

We formalize **agreement** or **consensus** in terms of distance between \(\pi_{1,y}\) and \(\pi_{2,y}\)
Formally
how does this relate to the standard framework?

We can still adopt the Predictive approach, as this it’s just another way of defining the summary statistic:

$$\rho_{\pi_A}(\theta|y_n)$$
Formally

how does this relate to the standard framework?

We can still adopt the Predictive approach, as this it’s just another way of defining the summary statistic:

$$\rho_{\pi_A}(\theta|y_n) \quad \rightarrow \quad d(\pi_1, y, \pi_2, y)$$
Formally
how does this relate to the standard framework?

We can still adopt the Predictive approach, as this it’s just another way of defining the summary statistic:

\[ \rho_{\pi_A}(\theta | y_n) \longrightarrow d(\pi_1, y, \pi_2, y) \]

\( PEC \) - Predictive Expectation Criterion

\[ e_{1,2}(n) = \mathbb{E}_{m_D}[d(\pi_1, y, \pi_2, y)] \]

\[ n^* = \min\{n \in \mathbb{N} : e_{1,2}(n) < \eta\} \]
Formally
how does this relate to the standard framework?

We can still adopt the Predictive approach, as this it’s just another way of defining the summary statistic:

\[ \rho_{\pi_A}(\theta | y_n) \rightarrow d(\pi_1, y, \pi_2, y) \]

- **PEC** - Predictive Expectation Criterion

\[ e_{1,2}(n) = \mathbb{E}_{m_D}[d(\pi_1, y, \pi_2, y)] \quad n^* = \min\{n \in \mathbb{N} : e_{1,2}(n) < \eta\} \]

- **PPC** - Predictive Probability Criterion

\[ p_{1,2}(n) = \mathbb{P}_{m_D}[d(\pi_1, y, \pi_2, y) > \gamma] \quad n^* = \min\{n \in \mathbb{N} : p_{1,2}(n) < \eta\} \]
Formally
how does this relate to the standard framework?

We can still adopt the Predictive approach, as this it’s just another way of defining the summary statistic:

\[ \rho_{\pi_A}(\theta|y_n) \quad \rightarrow \quad d(\pi_1, y, \pi_2, y) \]

- **PEC** - Predictive Expectation Criterion

\[ e_{1,2}(n) = \mathbb{E}_{m_D}[d(\pi_1, y, \pi_2, y)] \quad \quad n^* = \min\{n \in \mathbb{N} : e_{1,2}(n) < \eta\} \]

- **PPC** - Predictive Probability Criterion

\[ p_{1,2}(n) = \mathbb{P}_{m_D}[d(\pi_1, y, \pi_2, y) > \gamma] \quad \quad n^* = \min\{n \in \mathbb{N} : p_{1,2}(n) < \eta\} \]

we just need to pick a distance
Wasserstein distance
a.k.a. Kantorovic, Earth Mover

\((p, d)\) – Wasserstein distance

\(X \sim P\) and \(Y \sim Q\), \(p \geq 1\) and \(d\) ground distance

\[ W_{d,p}(P, Q) = \left( \inf_{J} \int_{\mathcal{X} \times \mathcal{Y}} d(x, y)^p \ dJ(x, y) \right)^{1/p} \]

where the infimum is over all joint distributions \(J\) having \(P\) and \(Q\) as marginals.
Wasserstein distance
it's this popular for a good reason

- it “metricises” convergence in distribution
  if two distributions are close w.r.t. the Wasserstein distance they are probabilistically similar.

Wasserstein distance
it's this popular for a good reason

> it “metricises” convergence in distribution
  if two distributions are close w.r.t. the Wasserstein distance they are probabilistically similar.

> it tells us why the distributions differ
  with the Wasserstein distance is associated to a map (transport plan) that shows us how we have to move the mass of $P$ to morph it into $Q$.

Wasserstein distance
it’s this popular for a good reason

- it “metricises” convergence in distribution
  if two distributions are close w.r.t. the Wasserstein distance they are
  probabilistically similar.

- it tells us why the distributions differ
  with the Wasserstein distance is associated to a map (transport plan) that
  shows us how we have to move the mass of $P$ to morph it into $Q$.

- it is sensible to the geometry of the space
  it’s not just about the location!

Let $X \sim N(\mu_X, \Sigma_X)$ and $Y \sim N(\mu_Y, \Sigma_Y)$, when the ground distance is taken to be the $L_2$ distance, we have a closed form expression for Wasserstein:

$$W_{L^2, 2}(X, Y) = \|\mu_X - \mu_Y\|^2_2 + B^2(\Sigma_X, \Sigma_Y)$$
Let $X \sim N(\mu_X, \Sigma_X)$ and $Y \sim N(\mu_Y, \Sigma_Y)$, when the ground distance is taken to be the $L_2$ distance, we have a closed form expression for Wasserstein:

$$W_{L^2}(X, Y) = \|\mu_X - \mu_Y\|_2^2 + B^2(\Sigma_X, \Sigma_Y)$$

$$B^2(\Sigma_X, \Sigma_Y) = \text{tr} \left[ \Sigma_X + \Sigma_Y - 2 \sqrt{\Sigma_X^{1/2} \Sigma_Y \Sigma_X^{1/2}} \right]$$

is the Bures distance.
Let \( X \sim N(\mu_X, \Sigma_X) \) and \( Y \sim N(\mu_Y, \Sigma_Y) \), when the ground distance is taken to be the \( L_2 \) distance, we have a closed form expression for Wasserstein:

\[
W_{L_2}(X, Y) = \|\mu_X - \mu_Y\|_2^2 + B^2(\Sigma_X, \Sigma_Y)
\]

\[
B^2(\Sigma_X, \Sigma_Y) = \text{tr} \left[ \Sigma_X + \Sigma_Y - 2 \sqrt{\Sigma_X^{1/2} \Sigma_Y \Sigma_X^{1/2}} \right]
\]

is the Bures distance.

distance between the means \( \star \) distance between the variances
Conjugate Univariate Gaussian Model
computing the Wasserstein distance

Likelihood: $N(\theta, \sigma^2)$, with $\sigma^2$ known.

$$\pi(\theta) = N \left( \theta; \mu_0, \frac{\sigma^2}{n_0} \right)$$
Conjugate Univariate Gaussian Model
computing the Wasserstein distance

**Likelihood:** \( N(\theta, \sigma^2) \), with \( \sigma^2 \) known.

\[
\pi(\theta) = N \left( \theta; \mu_o, \frac{\sigma^2}{n_o} \right) \quad \quad \quad \pi(\theta|y_n) = N \left( \theta; \frac{n_o \mu_o + n \bar{y}_n}{n + n_o}, \frac{\sigma^2}{n + n_o} \right)
\]
Conjugate Univariate Gaussian Model
computing the Wasserstein distance

**Likelihood:** \( N(\theta, \sigma^2) \), with \( \sigma^2 \) known.

\[
\pi(\theta) = N\left( \theta; \mu_0, \frac{\sigma^2}{n_0} \right) \quad \quad \pi(\theta|y_n) = N\left( \theta; \frac{n_0\mu_0 + n\bar{y}_n}{n + n_0}, \frac{\sigma^2}{n + n_0} \right)
\]

If we have two priors, Wasserstein between the corresponding posteriors is:

\[
W_{L^2,2}(\pi_1,y, \pi_2,y) = (\mu_{1,p} - \mu_{2,p})^2 + (\sigma_{1,p} - \sigma_{2,p})^2.
\]
Conjugate Gaussian Model
in the Bayesian predictive approach to SSD

Under the usual $\pi_D(\theta) \sim N(\mu_D, \sigma/n_D)$ assumption:
Under the usual $\pi_D(\theta) \sim N(\mu_D, \sigma/n_D)$ assumption:

\[
\text{PEC: } e_{1,2}(n) = \tilde{\mu}^2 + \sigma^2 \left( \frac{1}{n} + \frac{1}{n_D} \right) + \left( \frac{1}{\sqrt{n + n_1}} - \frac{1}{\sqrt{n + n_2}} \right)^2
\]
Conjugate Gaussian Model
in the Bayesian predictive approach to SSD

Under the usual $\pi_D(\theta) \sim N(\mu_D, \sigma/n_D)$ assumption:

**PEC:**

$$e_{1,2}(n) = \tilde{\mu}^2 + \sigma^2 \left( w_n^2 \left[ \frac{1}{n} + \frac{1}{n_D} \right] + \left[ \frac{1}{\sqrt{n + n_1}} - \frac{1}{\sqrt{n + n_2}} \right]^2 \right)$$

**PPC:**

$$p_{1,2}(n) = 1 - F_{\chi^2} \left( \frac{\gamma - B_{\sigma^2}}{\tilde{\sigma}^2}; df = 1, \tilde{\mu}^2 \right)$$

> $w_1 = n_1/(n + n_1)$
> $w_2 = n_2/(n + n_2)$
> $w_n = (1 - w_1) - (1 - w_2)$
> $\tilde{\mu} = w_1 \mu_1 - w_2 \mu_2 + w_n \mu_D$
> $\tilde{\sigma}^2 = w_n^2 \sigma^2 \left( \frac{1}{n} + \frac{1}{n_D} \right)$
> $B_{\sigma^2} = (\sigma_{1,p} - \sigma_{1,p})^2$
A Toy Example
mildly informative priors

\[ \pi_1(\theta) = N(0, 2/80) \quad \pi_2(\theta) = N(2, 2/50) \]
A Toy Example
mildly informative priors

\[ \eta = 0.1 \quad n^* = 125 \]
A Toy Example
mildly informative priors

\[ \eta = 0.1 \quad n^* = 125 \]
Another Toy Example

weakly informative priors

\[ \pi_1(\theta) = N(0, 2/8) \quad \pi_2(\theta) = N(2, 2/5) \]
Another Toy Example

weakly informative priors

\[ \eta = 0.1 \quad \text{and} \quad n^* = 15 \]
Another Toy Example
weakly informative priors

$$\eta = 0.1 \quad n^* = 15$$
How to select $\eta$?

a small bump in the road

Given a $\beta \in (0, 1)$, choose $\eta$ as

$$\beta \times \arg\max_n e_{1,2}(n)$$
How to select $\eta$?
a small bump in the road

Given a $\beta \in (0, 1)$, choose $\eta$ as

$$\beta \times \arg \max_n e_{1,2}(n)$$

It turns out that under some regularity assumptions, $e_{1,2}(n)$ can be monotone in $n$. 
How to select $\eta$?

Given a $\beta \in (0, 1)$, choose $\eta$ as

$$\beta \times \arg \max_n e_{1,2}(n)$$

It turns out that under some regularity assumptions, $e_{1,2}(n)$ can be monotone in $n$.

When this happens

$$\arg \max_n e_{1,2}(n) = e_{1,2}(1)$$

$\beta$ represent how much difference we can tolerate with respect to the minimum sample size possible.
A Toy Example

Reprise

\[ \eta = 0.1 \quad \text{and} \quad n^* = 125 \]
A Real Data Example
from Spiegelhalter et al. (2004)

\[ \theta = \log \text{OR of intravenous magnesium sulphate after acute myocardial infarction with respect to placebo.} \]
A Real Data Example
from Spiegelhlter et al. (2004)

\[ \theta = \log \text{OR of intravenous magnesium sulphate after acute myocardial infarction with respect to placebo.} \]

A bunch of priors encoding evidence from previous experiments:
An Unfair Comparison
was this really necessary?

- **Likelihood** Gaussian with unknown mean $\theta$ and $\sigma^2 = 4$
- **Design Prior** Gaussian with mean $\mu_D = 0.058$ and variance $\sigma^2/n_D$
- **Threshold** $\eta = 0.05$
An Unfair Comparison
was this really necessary?

- **Likelihood** Gaussian with unknown mean $\theta$ and $\sigma^2 = 4$
- **Design Prior** Gaussian with mean $\mu_D = 0.058$ and variance $\sigma^2 / n_D$
- **Threshold** $\eta = 0.05$

<table>
<thead>
<tr>
<th>$n_D$</th>
<th>$n^*_{\text{WASS}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4319</td>
<td>361</td>
</tr>
<tr>
<td>432</td>
<td>371</td>
</tr>
<tr>
<td>43</td>
<td>468</td>
</tr>
</tbody>
</table>
An Unfair Comparison
was this really necessary?

- **Likelihood** Gaussian with unknown mean $\theta$ and $\sigma^2 = 4$
- **Design Prior** Gaussian with mean $\mu_D = 0.058$ and variance $\sigma^2/n_D$
- **Threshold** $\eta = 0.05$

<table>
<thead>
<tr>
<th>$n_D$</th>
<th>$n_{WASS}^*$</th>
<th>$n_{MIXT}^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4319</td>
<td>361</td>
<td>498</td>
</tr>
<tr>
<td>432</td>
<td>371</td>
<td>509</td>
</tr>
<tr>
<td>43</td>
<td>468</td>
<td>190</td>
</tr>
</tbody>
</table>

Consensus does not typically come “for free”

When the posterior distributions are not Gaussian, the Wasserstein distance does not necessarily have an analytic expression.

This is the case for the Beta-Binomial conjugate model.
Conjugate Beta-Binomial
moving beyond gaussianity

When the posterior distributions are not Gaussian, the Wasserstein distance does not necessarily have an analytic expression.

This is the case for the Beta-Binomial conjugate model.

Possible solutions are:
  > Numerical evaluation of the Wasserstein distance
  > Approximation of the Wasserstein distance via Stein’s method
X, Y random variables (typically X is "what you have", Y is "what you want")

1. rewrite the distance between X and Y as the expectation of a functional $h(X)$
X, Y random variables (typically X is "what you have", Y is "what you want")

1. rewrite the distance between X and Y as the expectation of a functional $h(X)$
Stein’s method
quickest introduction ever

X, Y random variables (typically X is "what you have", Y is "what you want")

1. rewrite the distance between X and Y as the expectation of a functional $h(X)$

2. **bound** such expectation
Stein’s method
quickest introduction ever

X, Y random variables (typically X is "what you have", Y is "what you want")

1. rewrite the distance between X and Y as the expectation of a functional $h(X)$

2. bound such expectation

X, Y random variables (typically X is "what you have", Y is "what you want")

1. rewrite the distance between X and Y as the **expectation** of a functional $h(X)$

2. **bound** such expectation

If we compare X and Y via the $L_1$ Wasserstein distance, we can derive tight bounds for it.

---

Stein’s bound for the B-B case
the second most famous framework in clinical trials

**Likelihood:** Binomial(θ, N), T events in the sample.

\[ \pi(\theta) = \text{Beta}(\theta; \alpha, \beta) \]
Stein’s bound for the B-B case
the second most famous framework in clinical trials

**Likelihood:** Binomial($\theta, N$), $T$ events in the sample.

\[
\pi(\theta) = \text{Beta}(\theta; \alpha, \beta) \quad \quad \pi(\theta|y_n) = \text{Beta}(\theta; \alpha + t, \beta + n - t)
\]
Stein’s bound for the B-B case
the second most famous framework in clinical trials

**Likelihood:** Binomial($\theta, N$), $T$ events in the sample.

$$
\pi(\theta) = \text{Beta}(\theta; \alpha, \beta) \\
\pi(\theta|y_n) = \text{Beta}(\theta; \alpha + t, \beta + n - t)
$$

$$
d_W(\pi_1, y, \pi_2, y) \leq \frac{|\alpha_1 - \alpha_2|}{\alpha_1 + \beta_1 + n} (1 - \mu_{2,p}) + \frac{|\beta_2 - \beta_1|}{\alpha_1 + \beta_1 + n} \mu_{2,p}
$$
Stein’s bound for the B-B case
the second most famous framework in clinical trials

**Likelihood:** Binomial(θ, N), T events in the sample.

\[ \pi(\theta) = \text{Beta}(\theta; \alpha, \beta) \quad \pi(\theta|y_n) = \text{Beta}(\theta; \alpha + t, \beta + n - t) \]

\[ d_W(\pi_1, y, \pi_2, y) \leq \frac{|\alpha_1 - \alpha_2|}{\alpha_1 + \beta_1 + n} (1 - \mu_{2,p}) + \frac{|\beta_2 - \beta_1|}{\alpha_1 + \beta_1 + n} \mu_{2,p} \]

If we assume \( \pi_D(\theta) = \text{Beta}(\theta; \alpha_D, \beta_D) \) it is possible to bound the PEC and PPC just by remembering:

\[ \mathbb{E}_{m_D}[T] = \frac{n\alpha_D}{\alpha_D + \beta_D} \]
Yet Another Toy Example

the more the merrier

\[ \pi_1(\theta) = \text{Beta}(9, 13) \]
\[ \pi_2(\theta) = \text{Beta}(12, 4) \]
Yet Another Toy Example
the more the merrier

\[ \eta = 0.1 \quad \text{and} \quad n^* = 184 \]
R-package coming soon!

tulliapadellini.github.io
tullia.padellini@uniroma1.it

Thanks!