Sequential Stopping for High-Throughput Experiments

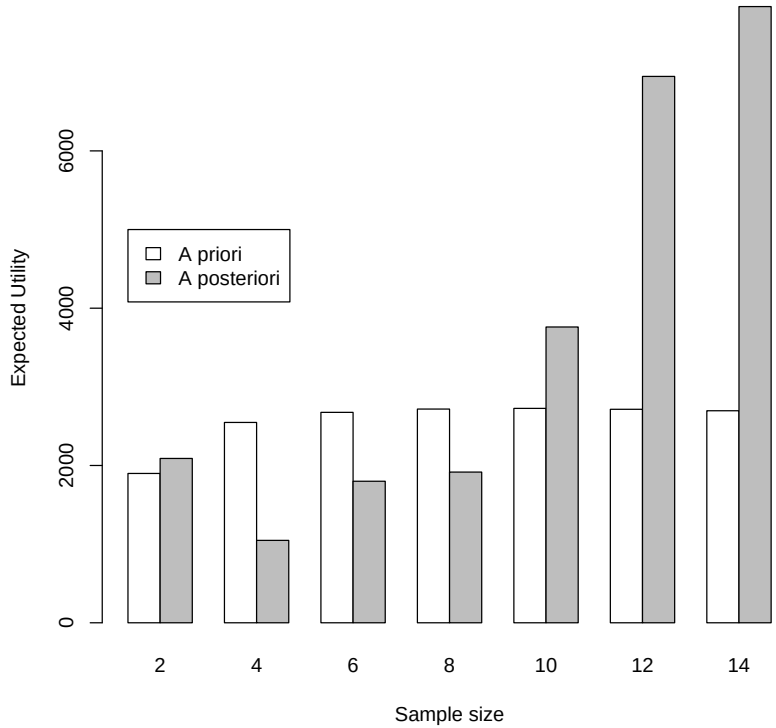
David Rossell[1] and Peter Müller[2]

[1]Biostatistics and Bioinformatics Unit, Institute for Research in Biomedicine of Barcelona, Spain; [2]UT Austin

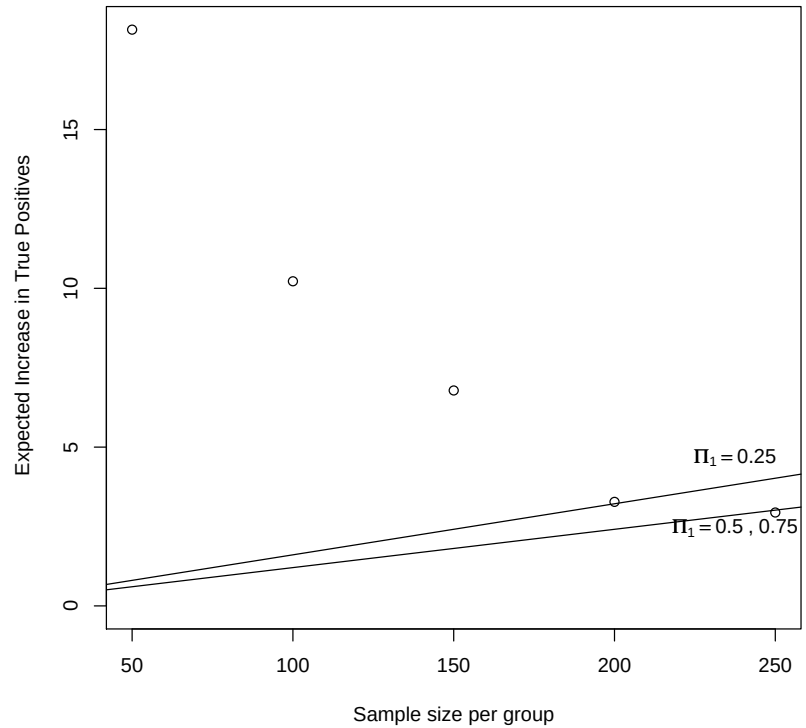
August 10, 2011

Expected Utility vs. Sample Size

Determine sample size by utility maximization.



prior expected utility at $t = 0$ (white)
for fixed sample size
vs. post exp utility ($t = \tau$) (grey)
for sequential stopping



optimal decision boundaries
for sequential stopping

Setup

Outcome: x_{ij} outcome for gene (protein, ...) i, $i = 1, \ldots, n$ and sample j, $j = 1, \ldots, T$

Setup

Outcome: x_{ij} outcome for gene (protein, ...) i, $i = 1, \ldots, n$ and sample j, $j = 1, \ldots, T$

Condition: $z_j \in \{0, \ldots, n_z\}$, biologic condition (cancer type, normal etc.)

Setup

Outcome: x_{ij} outcome for gene (protein, ...) i, $i = 1, \ldots, n$ and sample j, $j = 1, \ldots, T$

Condition: $z_j \in \{0, \ldots, n_z\}$, biologic condition (cancer type, normal etc.)

Differential expression: $\delta_i \in \{0, 1\}$, unknown truth (parameter)

Setup

Outcome: x_{ij} outcome for gene (protein, ...) i, $i = 1, \ldots, n$ and sample j, $j = 1, \ldots, T$

Condition: $z_j \in \{0, \ldots, n_z\}$, biologic condition (cancer type, normal etc.)

Differential expression: $\delta_i \in \{0, 1\}$, unknown truth (parameter)

Parameters: θ_i and hyperpars ν, indexes sampling model $p(x_{i1}, \ldots, x_{iT} \mid \theta_i, \nu)$.

Setup

Outcome: x_{ij} outcome for gene (protein, ...) i, $i = 1, \ldots, n$ and sample j, $j = 1, \ldots, T$

Condition: $z_j \in \{0, \ldots, n_z\}$, biologic condition (cancer type, normal etc.)

Differential expression: $\delta_i \in \{0, 1\}$, unknown truth (parameter)

Parameters: θ_i and hyperpars ν, indexes sampling model $p(x_{i1}, \ldots, x_{iT} \mid \theta_i, \nu)$.

Data at time t: $x_t = \{x_{it}, 1 \leq i \leq n\}$
$x_{1:t} = \{x_{ij}, 1 \leq i \leq n,\ 1 \leq j \leq t\}$

Model

Any!

Any! Well, need quick & easy $p(\delta_j \mid x_{1:t})$.

Any! Well, need quick & easy $p(\delta_j \mid x_{1:t})$.

GaGa model: Rossell (2009 AOAS), generalizes Ga/Ga hierarchical model of Kendziorski et al. (2003 Biostat) with gen-specific CV.

Any! Well, need quick & easy $p(\delta_j \mid x_{1:t})$.

GaGa model: Rossell (2009 AOAS), generalizes Ga/Ga hierarchical model of Kendziorski et al. (2003 Biostat) with gen-specific CV.

Sampling model: $x_{ij} \sim \text{Ga}(\alpha_i, \alpha_i/\lambda_{iz_j})$ with $E(x_{ij}) = \lambda_{iz}$ and $CV = 1/\sqrt{\alpha_i}$.

Any! Well, need quick & easy $p(\delta_j \mid x_{1:t})$.

GaGa model: Rossell (2009 AOAS), generalizes Ga/Ga hierarchical model of Kendziorski et al. (2003 Biostat) with gen-specific CV.

Sampling model: $x_{ij} \sim \text{Ga}(\alpha_i, \alpha_i/\lambda_{iz_j})$ with $E(x_{ij}) = \lambda_{iz}$ and $CV = 1/\sqrt{\alpha_i}$.

Gene-specific pars $\theta_i = (\lambda_{i0}, \lambda_{i1}, \alpha_i)$

Any! Well, need quick & easy $p(\delta_j \mid x_{1:t})$.

GaGa model: Rossell (2009 AOAS), generalizes Ga/Ga hierarchical model of Kendziorski et al. (2003 Biostat) with gen-specific CV.

Sampling model: $x_{ij} \sim \text{Ga}(\alpha_i, \alpha_i/\lambda_{iz_j})$ with $E(x_{ij}) = \lambda_{iz}$ and $CV = 1/\sqrt{\alpha_i}$.

Gene-specific pars $\theta_i = (\lambda_{i0}, \lambda_{i1}, \alpha_i)$

Prior: allows gene-specific CV & $p(\lambda_{i0} = \lambda_{i1}) > 0$

- conjugate IG prior $\lambda_{i0}^{-1} \sim \text{Ga}(\alpha_0, \alpha_0/\nu)$
- pos prob for $\delta_i = I(\lambda_{i0} = \lambda_{i1})$, i.e., non-differential expression.
- $\alpha_i|\beta, \mu \sim \text{Ga}(\beta, \beta/\mu)$

Any! Well, need quick & easy $p(\delta_j \mid x_{1:t})$.

GaGa model: Rossell (2009 AOAS), generalizes Ga/Ga hierarchical model of Kendziorski et al. (2003 Biostat) with gen-specific CV.

Sampling model: $x_{ij} \sim \text{Ga}(\alpha_i, \alpha_i/\lambda_{iz_j})$ with $E(x_{ij}) = \lambda_{iz}$ and $CV = 1/\sqrt{\alpha_i}$.

Gene-specific pars $\theta_i = (\lambda_{i0}, \lambda_{i1}, \alpha_i)$

Prior: allows gene-specific CV & $p(\lambda_{i0} = \lambda_{i1}) > 0$

- conjugate IG prior $\lambda_{i0}^{-1} \sim \text{Ga}(\alpha_0, \alpha_0/\nu)$
- pos prob for $\delta_i = I(\lambda_{i0} = \lambda_{i1})$, i.e., non-differential expression.
- $\alpha_i|\beta, \mu \sim \text{Ga}(\beta, \beta/\mu)$

Posterior: can analytically marginalize over $(\lambda_{i0}, \lambda_{i1}, \alpha_i)$.

Any! Well, need quick & easy $p(\delta_j \mid x_{1:t})$.

GaGa model: Rossell (2009 AOAS), generalizes Ga/Ga hierarchical model of Kendziorski et al. (2003 Biostat) with gen-specific CV.

Sampling model: $x_{ij} \sim \text{Ga}(\alpha_i, \alpha_i/\lambda_{iz_j})$ with $E(x_{ij}) = \lambda_{iz}$ and $CV = 1/\sqrt{\alpha_i}$.

Gene-specific pars $\theta_i = (\lambda_{i0}, \lambda_{i1}, \alpha_i)$

Prior: allows gene-specific CV & $p(\lambda_{i0} = \lambda_{i1}) > 0$

- conjugate IG prior $\lambda_{i0}^{-1} \sim \text{Ga}(\alpha_0, \alpha_0/\nu)$
- pos prob for $\delta_i = I(\lambda_{i0} = \lambda_{i1})$, i.e., non-differential expression.
- $\alpha_i|\beta, \mu \sim \text{Ga}(\beta, \beta/\mu)$

Posterior: can analytically marginalize over $(\lambda_{i0}, \lambda_{i1}, \alpha_i)$.

Prob of DE: main summary of interst

$$p_i \equiv p(\delta_i = 1 \mid \mathbf{x}_{1:T})$$

Decisions: two decisions

- sequential stopping decision, $s_i = s(\mathbf{x}_{1:t}) = 1$ (stopping) vs. $s_i = 0$ (continuation)

Decisions: two decisions

- sequential stopping decision, $s_i = s(\mathbf{x}_{1:t}) = 1$ (stopping) vs. $s_i = 0$ (continuation)
  Alternatively use $\tau = \min\{t : \ s_t = 1\}$

Decisions: two decisions

- sequential stopping decision, $s_i = s(\mathbf{x}_{1:t}) = 1$ (stopping) vs. $s_i = 0$ (continuation)
  Alternatively use $\tau = \min\{t : \ s_t = 1\}$
- terminal decision, $\mathbf{d}(\mathbf{x}_{1:t} = (d_1(\mathbf{x}_{1:t}), \ldots, d_n(\mathbf{x}_{1:t}))$
  $d_i(\cdot) = 1$ for flagging gene i.

Optimal Stopping

Decisions: two decisions

- sequential stopping decision, $s_i = s(\mathbf{x}_{1:t}) = 1$ (stopping) vs. $s_i = 0$ (continuation)
  Alternatively use $\tau = \min\{t : \ s_t = 1\}$
- terminal decision, $\mathbf{d}(\mathbf{x}_{1:t} = (d_1(\mathbf{x}_{1:t}), \ldots, d_n(\mathbf{x}_{1:t}))$
  $d_i(\cdot) = 1$ for flagging gene i.

Bayes rule: optimal decision w.r.t. a utility function

$$u(d, s, y, \theta)$$

Optimal Stopping

Decisions: two decisions

- sequential stopping decision, $s_i = s(\mathbf{x}_{1:t}) = 1$ (stopping) vs. $s_i = 0$ (continuation)
  Alternatively use $\tau = \min\{t : \ s_t = 1\}$
- terminal decision, $\mathbf{d}(\mathbf{x}_{1:t} = (d_1(\mathbf{x}_{1:t}), \ldots, d_n(\mathbf{x}_{1:t}))$
  $d_i(\cdot) = 1$ for flagging gene i.

Bayes rule: optimal decision w.r.t. a utility function

$$u(d, s, y, \theta)$$

quantifies preferences for decision (d, s) under future data $\mathbf{x}$ and hypothetical truth θ.

Optimal Stopping

Decisions: two decisions

- sequential stopping decision, $s_i = s(\mathbf{x}_{1:t}) = 1$ (stopping) vs. $s_i = 0$ (continuation)
  Alternatively use $\tau = \min\{t : \ s_t = 1\}$
- terminal decision, $\mathbf{d}(\mathbf{x}_{1:t} = (d_1(\mathbf{x}_{1:t}), \ldots, d_n(\mathbf{x}_{1:t}))$
  $d_i(\cdot) = 1$ for flagging gene i.

Bayes rule: optimal decision w.r.t. a utility function

$$u(d, s, y, \theta)$$

quantifies preferences for decision (d, s) under future data $\mathbf{x}$ and hypothetical truth θ.

$$d^\star(\mathbf{x}) = \arg\max \int u(d, s, \mathbf{x}, \theta)\, dp(\theta \mid \mathbf{x})$$

Optimal Stopping

Decisions: two decisions

- sequential stopping decision, $s_i = s(\mathbf{x}_{1:t}) = 1$ (stopping) vs. $s_i = 0$ (continuation)
  Alternatively use $\tau = \min\{t : \ s_t = 1\}$
- terminal decision, $\mathbf{d}(\mathbf{x}_{1:t} = (d_1(\mathbf{x}_{1:t}), \ldots, d_n(\mathbf{x}_{1:t}))$
  $d_i(\cdot) = 1$ for flagging gene i.

Bayes rule: optimal decision w.r.t. a utility function

$$u(d, s, y, \theta)$$

quantifies preferences for decision (d, s) under future data $\mathbf{x}$ and hypothetical truth θ.

$$d^\star(\mathbf{x}) = \arg\max \int u(d, s, \mathbf{x}, \theta)\, dp(\theta \mid \mathbf{x})$$

and a bit more complicated for the sequential decision s

Utility

Utility: use two different criteria for stopping and terminal decision (not quite kosher..)

- Terminal decision: Bayes rule under FDR control. Let $D = \sum d_i$, number of $\widehat{\text{DE}}$ genes

$$\text{FDR} = \sum d_i(\mathbf{x}_{1:t})(1 - \delta_i)/D$$

Report max D subject to $E(\text{FDR} \mid data) \leq k$

Utility

Utility: use two different criteria for stopping and terminal decision (not quite kosher..)

- Terminal decision: Bayes rule under FDR control. Let $D = \sum d_i$, number of $\widehat{\text{DE}}$ genes

$$\text{FDR} = \sum d_i(\mathbf{x}_{1:t})(1 - \delta_i)/D$$

  Report max D subject to $E(\text{FDR} \mid data) \leq k$

- Stopping decision

$$u(s_t = 1, \mathbf{d}(\mathbf{x}_{1:\tau}), \mathbf{x}_{1:\tau}, \boldsymbol{\delta},) = -c\tau + \sum_{i=1}^{n} \delta_i d_i(\mathbf{x}_{1:\tau}).$$

  c : # of true pos that justify one more obs.

Expected Utility & Optimal Decision

Terminal decision: maximize utility in expectation w.r.t. δ, conditional on $\mathbf{x}_{1:\tau}$.

Expected Utility & Optimal Decision

Terminal decision: maximize utility in expectation w.r.t. δ, conditional on $\mathbf{x}_{1:\tau}$.

$$d_i(\mathbf{x}_{1:\tau}) = I(p_i > \kappa)$$

Report genes with highest posterior prob of DE (not obvious, but simple algebra)

Expected Utility & Optimal Decision

Terminal decision: maximize utility in expectation w.r.t. δ, conditional on $\mathbf{x}_{1:\tau}$.

$$d_i(\mathbf{x}_{1:\tau}) = I(p_i > \kappa)$$

Report genes with highest posterior prob of DE (not obvious, but simple algebra)

Stopping decision: compare expected util for $s_t = 1$ vs. $s_t = 0$

Expected Utility & Optimal Decision

Terminal decision: maximize utility in expectation w.r.t. δ, conditional on $\mathbf{x}_{1:\tau}$.

$$d_i(\mathbf{x}_{1:\tau}) = I(p_i > \kappa)$$

Report genes with highest posterior prob of DE (not obvious, but simple algebra)

Stopping decision: compare expected util for $s_t = 1$ vs. $s_t = 0$

Expected utility: Easy for $s_t = 1$ (stopping)

$$\overline{u}_t(s_t = 1, \mathbf{x}_{1:t}) = -ct + \sum_{i=1}^{n} P(\delta_i = 1 \mid \mathbf{x}_{1:t}) d_i(\mathbf{x}_{1:t}),$$

Expected Utility & Optimal Decision

Terminal decision: maximize utility in expectation w.r.t. δ, conditional on $\mathbf{x}_{1:\tau}$.

$$d_i(\mathbf{x}_{1:\tau}) = I(p_i > \kappa)$$

Report genes with highest posterior prob of DE (not obvious, but simple algebra)

Stopping decision: compare expected util for $s_t = 1$ vs. $s_t = 0$

Expected utility: Easy for $s_t = 1$ (stopping)

$$\overline{u}_t(s_t = 1, \mathbf{x}_{1:t}) = -ct + \sum_{i=1}^{n} P(\delta_i = 1 \mid \mathbf{x}_{1:t}) d_i(\mathbf{x}_{1:t}),$$

Continuation:

$$\overline{u}_t(s_t = 0, \mathbf{x}_{1:t}) = -c + \int \overline{u}_t(s^{\star}_{t+1}, \mathbf{x}_{1:t}, x_{t+1})\, dp(x_{t+1} \mid \mathbf{x}_{1:t})$$

Expected Utility & Optimal Decision

Terminal decision: maximize utility in expectation w.r.t. δ, conditional on $\mathbf{x}_{1:\tau}$.

$$d_i(\mathbf{x}_{1:\tau}) = I(p_i > \kappa)$$

Report genes with highest posterior prob of DE (not obvious, but simple algebra)

Stopping decision: compare expected util for $s_t = 1$ vs. $s_t = 0$

Expected utility: Easy for $s_t = 1$ (stopping)

$$\overline{u}_t(s_t = 1, \mathbf{x}_{1:t}) = -ct + \sum_{i=1}^{n} P(\delta_i = 1 \mid \mathbf{x}_{1:t}) d_i(\mathbf{x}_{1:t}),$$

Continuation:

$$\overline{u}_t(s_t = 0, \mathbf{x}_{1:t}) = -c + \int \overline{u}_t(s^\star_{t+1}, \mathbf{x}_{1:t}, x_{t+1})\, dp(x_{t+1} \mid \mathbf{x}_{1:t})$$

requires integration w.r.t. x_{t+1} and plug-in of optimal rule $s^\star_{t+1}(\cdot)$ $\rightarrow$ backward induction (arrrgh!)

Approximate Solution

Brockwell & Kadanae (2003 JCGS); Müller et al. (2006 JSPI), Rossell et al. (2006 Biostat), and others:

- approximate seq design by decision boundaries on summary statistics.

Approximate Solution

Brockwell & Kadanae (2003 JCGS); Müller et al. (2006 JSPI), Rossell et al. (2006 Biostat), and others:

- approximate seq design by decision boundaries on summary statistics.
  - restricted decision rules facilitate backward induction.

Approximate Solution

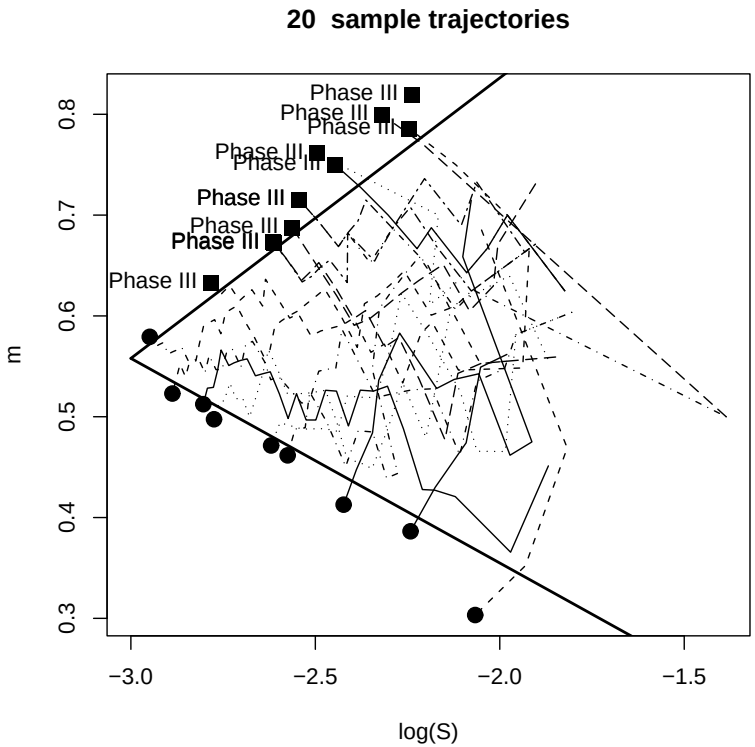
Brockwell & Kadanae (2003 JCGS); Müller et al. (2006 JSPI), Rossell et al. (2006 Biostat), and others:

- approximate seq design by decision boundaries on summary statistics.
  - restricted decision rules facilitate backward induction.
  - Forward simulation to evaluate required post pred expectations.

Approximate Solution

Brockwell & Kadanae (2003 JCGS); Müller et al. (2006 JSPI), Rossell et al. (2006 Biostat), and others:

- approximate seq design by decision boundaries on summary statistics.
  - restricted decision rules facilitate backward induction.
  - Forward simulation to evaluate required post pred expectations.

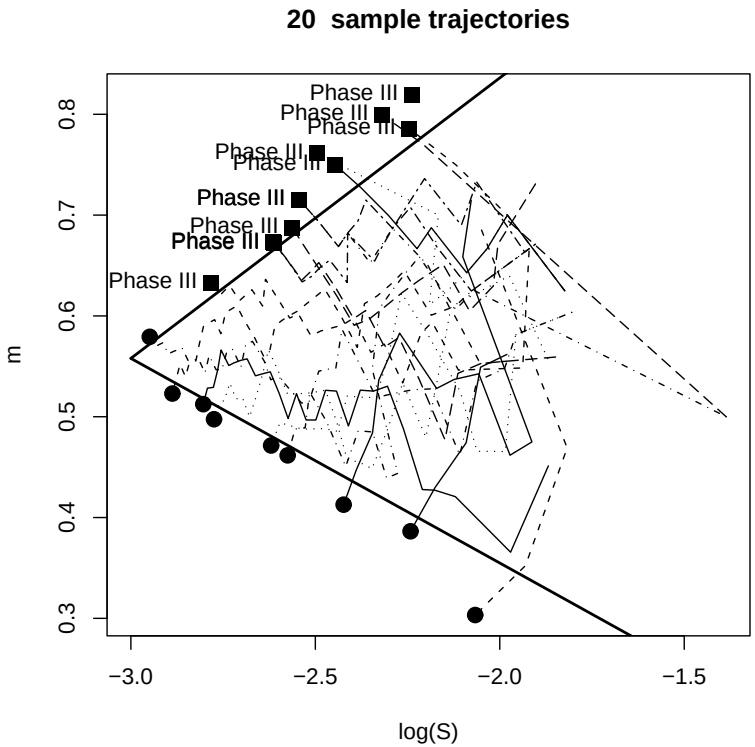


- Summary statistic $Y = (S, m)$ and decision boundaries.
  Each trajectory shows a possible trial realization.

Approximate Solution

Brockwell & Kadanae (2003 JCGS); Müller et al. (2006 JSPI), Rossell et al. (2006 Biostat), and others:

- approximate seq design by decision boundaries on summary statistics.
  - restricted decision rules facilitate backward induction.
  - Forward simulation to evaluate required post pred expectations.



- Summary statistic $Y = (S, m)$ and decision boundaries.
  Each trajectory shows a possible trial realization.
- When the summary Y crosses decision boundary the trial stops.

Summary Statistic

Increment in true pos:

$$\Delta_t(\text{TP}) \equiv E_{\mathbf{x}_{t+1}}\left[\overline{u}_{t+1}(s_{t+1}=1, \mathbf{x}_{1:t+1}) \mid \mathbf{x}_{1:t}\right] - \overline{u}_t(s_t = 1, \mathbf{x}_{1:t})$$

$E_{\mathbf{x}_{t+1}}(\cdot|\mathbf{x}_{1:t})$, conditional on $\mathbf{x}_{1:t}$ and marginalizing w.r.t. to future data $\mathbf{x}_{t+1}$ and θ.

Summary Statistic

Increment in true pos:

$$\Delta_t(\text{TP}) \equiv E_{\mathbf{x}_{t+1}}\left[\overline{u}_{t+1}(s_{t+1}=1, \mathbf{x}_{1:t+1}) \mid \mathbf{x}_{1:t}\right] - \overline{u}_t(s_t = 1, \mathbf{x}_{1:t})$$

$E_{\mathbf{x}_{t+1}}(\cdot|\mathbf{x}_{1:t})$, conditional on $\mathbf{x}_{1:t}$ and marginalizing w.r.t. to future data $\mathbf{x}_{t+1}$ and θ.

Summary: $Y = (t, \Delta_t(\text{TP}))$

Summary Statistic

Increment in true pos:

$$\Delta_t(\text{TP}) \equiv E_{\mathbf{x}_{t+1}}\left[\overline{u}_{t+1}(s_{t+1}=1, \mathbf{x}_{1:t+1}) \mid \mathbf{x}_{1:t}\right] - \overline{u}_t(s_t = 1, \mathbf{x}_{1:t})$$

$E_{\mathbf{x}_{t+1}}(\cdot|\mathbf{x}_{1:t})$, conditional on $\mathbf{x}_{1:t}$ and marginalizing w.r.t. to future data $\mathbf{x}_{t+1}$ and θ.

Summary: $Y = (t, \Delta_t(\text{TP}))$

Decision boundary: stop sampling when $(t, \Delta_t(\text{TP}))$ crosses a line.

Summary Statistic

Increment in true pos:

$$\Delta_t(\text{TP}) \equiv E_{\mathbf{x}_{t+1}}\left[\overline{u}_{t+1}(s_{t+1}=1, \mathbf{x}_{1:t+1}) \mid \mathbf{x}_{1:t}\right] - \overline{u}_t(s_t = 1, \mathbf{x}_{1:t})$$

$E_{\mathbf{x}_{t+1}}(\cdot|\mathbf{x}_{1:t})$, conditional on $\mathbf{x}_{1:t}$ and marginalizing w.r.t. to future data $\mathbf{x}_{t+1}$ and θ.

Summary: $Y = (t, \Delta_t(\text{TP}))$

Decision boundary: stop sampling when $(t, \Delta_t(\text{TP}))$ crosses a line.

$\rightarrow$ a slight generalization of traditional myopic design

Approx Optimal Stopping – Optimal Decision Boundary

Decision boundaries: parametrize boundary by β, e.g. lines.

Approx Optimal Stopping – Optimal Decision Boundary

Decision boundaries: parametrize boundary by β, e.g. lines.
Optimal boundary is a simple non-seq optimization problem.

Approx Optimal Stopping – Optimal Decision Boundary

Decision boundaries: parametrize boundary by β, e.g. lines.
Optimal boundary is a simple non-seq optimization problem.

Forward simulation: simulate possible study realizations $(x_{1:T}^{(b)})$, $b = 1, \ldots, B$

Approx Optimal Stopping – Optimal Decision Boundary

Decision boundaries: parametrize boundary by β, e.g. lines.
Optimal boundary is a simple non-seq optimization problem.

Forward simulation: simulate possible study realizations $(x_{1:T}^{(b)})$, $b = 1, \ldots, B$

Expected utility: Monte Carlo approx

$$\overline{U}(\beta) = 1/B \sum \overline{u}(s_{\tau^{(b)}} = 1, \mathbf{x}_{1:\tau}^{(b)})$$

Approx Optimal Stopping – Optimal Decision Boundary

Decision boundaries: parametrize boundary by β, e.g. lines.
Optimal boundary is a simple non-seq optimization problem.

Forward simulation: simulate possible study realizations $(x_{1:T}^{(b)})$, $b = 1, \ldots, B$

Expected utility: Monte Carlo approx

$$\overline{U}(\beta) = 1/B \sum \overline{u}(s_{\tau(b)} = 1, \mathbf{x}_{1:\tau}^{(b)})$$

Optimal decision: optimize $\overline{U}(\beta)$ over β.

Example 1: Next Generation Sequencing

Pilot data: RNA-sequencing.

2 human muscle & 1 human brain, compare gene expression.

Design: up to $T = 5$ more samples per group.

Example 1: Next Generation Sequencing

Pilot data: RNA-sequencing.

2 human muscle & 1 human brain, compare gene expression.

Design: up to $T = 5$ more samples per group.

Optimal design: using decision boundaries and post pred given the pilot data. $\rightarrow$ next slide

Example 1: Next Generation Sequencing

Pilot data: RNA-sequencing.

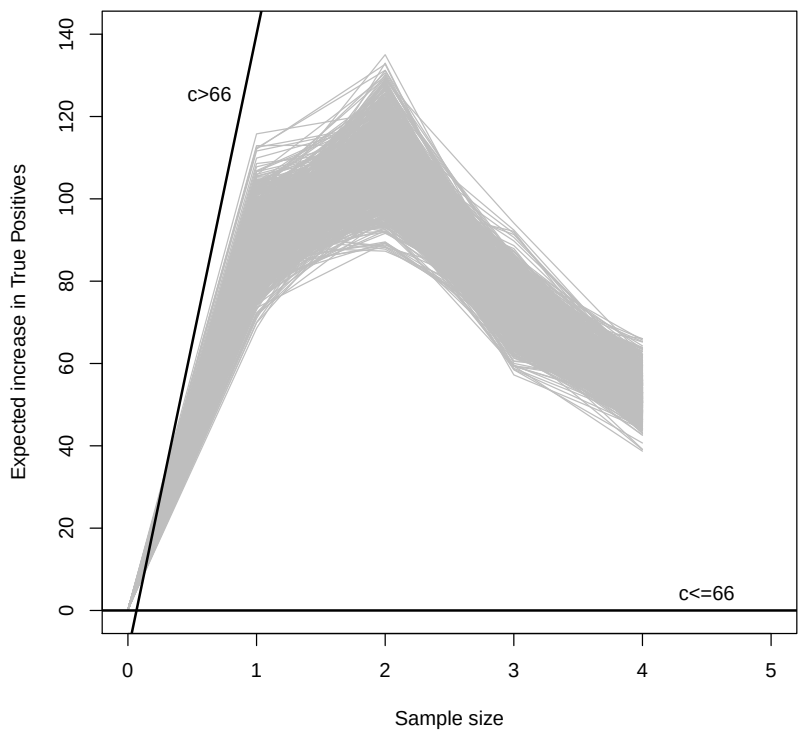
2 human muscle & 1 human brain, compare gene expression.

Design: up to $T = 5$ more samples per group.

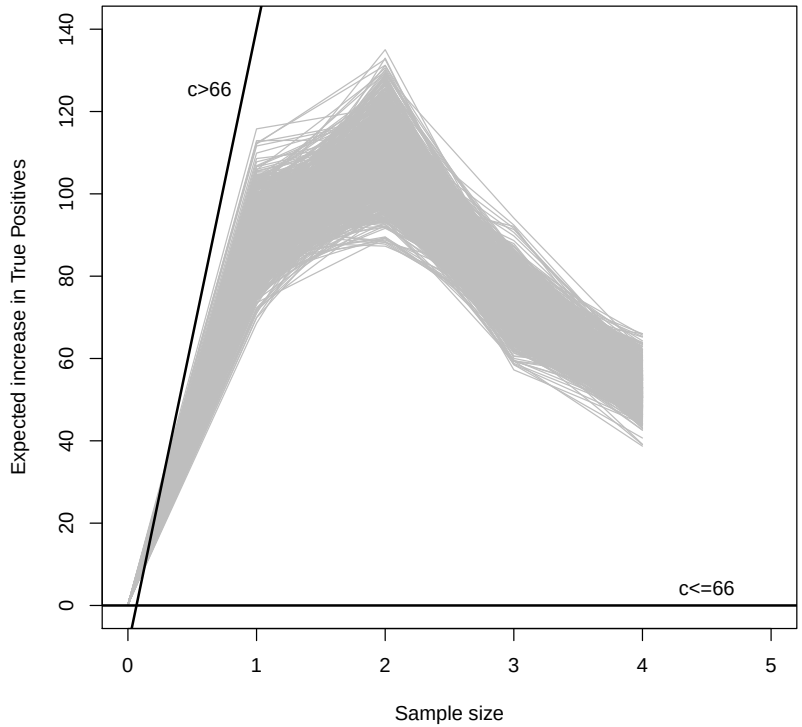
Optimal design: using decision boundaries and post pred given the pilot data. $\rightarrow$ next slide

GaGa model for $x_{ij} = \log(RPKM + 1)$.

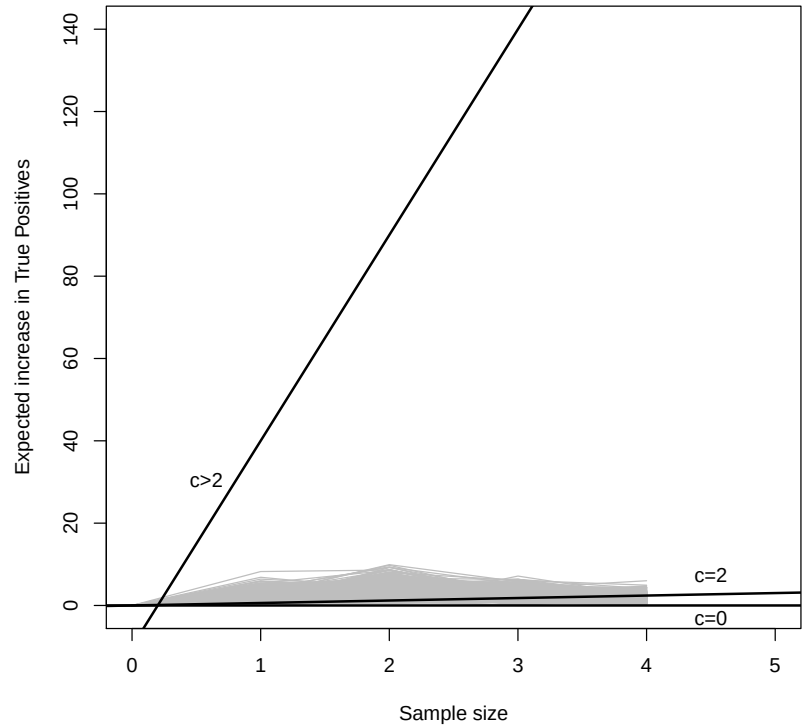
Results


c>66
c<=66
Expected increase in True Positives
Sample size
0
20
40
60
80
100
120
140
0
1
2
3
4
5

Results



(a) brain vs. muscle
$s_t = 0$
for most posterior pred simulations



(b) muscle vs. muscle
$s_t = 1$
(unless $c < 2$)

Example 2: Microarray Data

Study: Campo Dell'Orto et al. (2007), leukemia microarray; 21 ALL & 15 MLL patients, 54,675 genes, Affymetrix array

Design: collect data in batches of 2 arrays each per group, max of $T = 7$ batches.

Example 2: Microarray Data

Study: Campo Dell'Orto et al. (2007), leukemia microarray; 21 ALL & 15 MLL patients, 54,675 genes, Affymetrix array

Design: collect data in batches of 2 arrays each per group, max of $T = 7$ batches.

Hyperpars: estimate hyperparameters from Armstrong et al. (2002) data (same arrays, similar patient populations)

Example 2: Microarray Data

Study: Campo Dell'Orto et al. (2007), leukemia microarray; 21 ALL & 15 MLL patients, 54,675 genes, Affymetrix array

Design: collect data in batches of 2 arrays each per group, max of $T = 7$ batches.

Hyperpars: estimate hyperparameters from Armstrong et al. (2002) data (same arrays, similar patient populations)

Results: based on post pred given Armstrong data $\rightarrow$ next slide

Example 2: Microarray Data

Study: Campo Dell'Orto et al. (2007), leukemia microarray; 21 ALL & 15 MLL patients, 54,675 genes, Affymetrix array

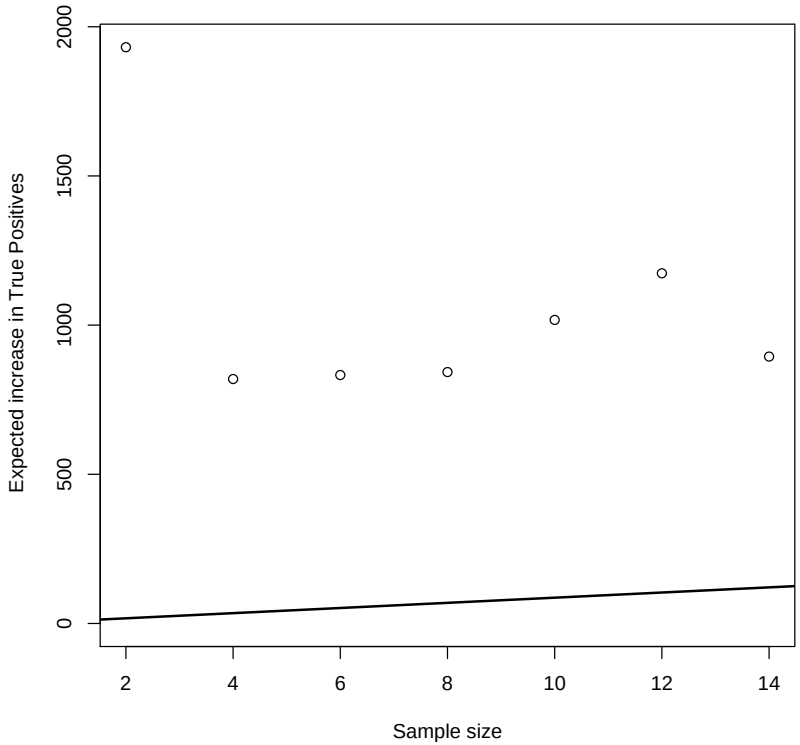
Design: collect data in batches of 2 arrays each per group, max of $T = 7$ batches.

Hyperpars: estimate hyperparameters from Armstrong et al. (2002) data (same arrays, similar patient populations)

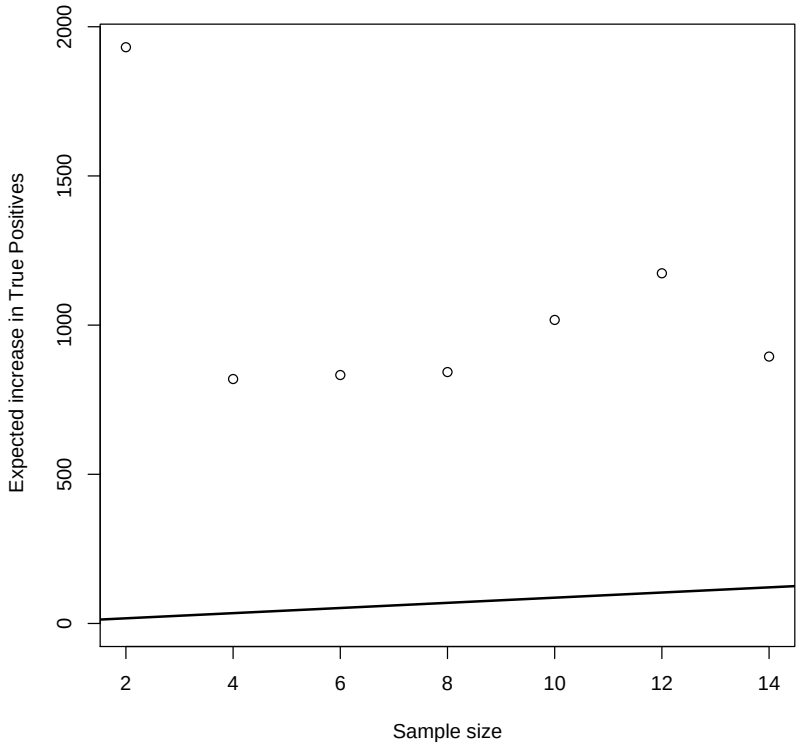
Results: based on post pred given Armstrong data $\rightarrow$ next slide

GaGa model on x_{ij} = GCRMA pre-processed scores (Wu & Irizarry, 2007).

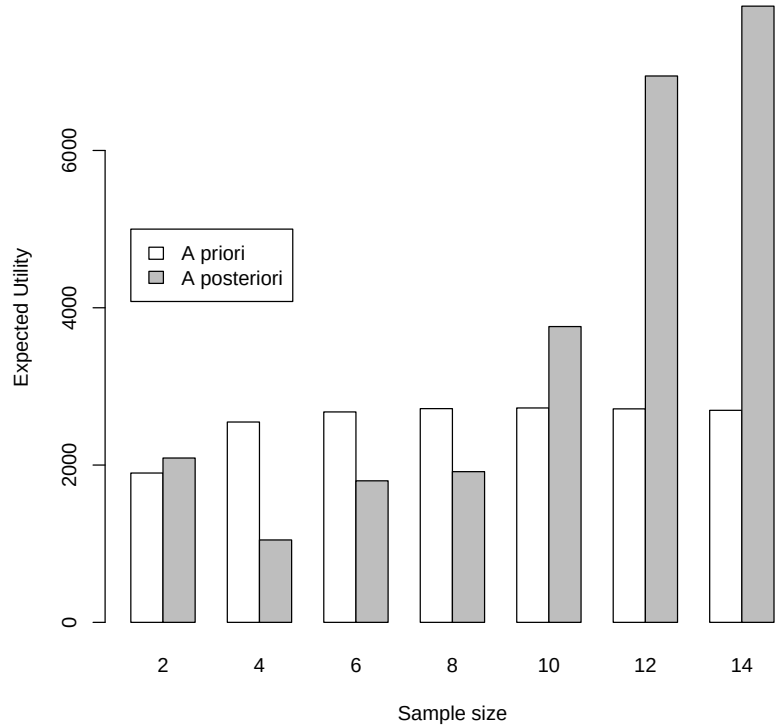
Results


Expected increase in True Positives
0
500
1000
1500
2000
2
4
6
8
10
12
14
Sample size

Results



optimal decision boundaries



fixed sample size design (prior) (white)
seq stopping
(posterior at $t = \tau$) (grey)

Example 3: RPPA

Study: Korenblau et al. (2006), protein abundance in bone marrow (BMA) vs. peripheral blood (PB), using RPPA

Data: 168 proteins, 435 BMA + 282 PB samples.

Design: batches of 50 samples per group, max of $T = 250$. RPPA is cheap $\rightarrow$ $c = 5$ or 10.

Example 3: RPPA

Study: Korenblau et al. (2006), protein abundance in bone marrow (BMA) vs. peripheral blood (PB), using RPPA

Data: 168 proteins, 435 BMA + 282 PB samples.

Design: batches of 50 samples per group, max of $T = 250$. RPPA is cheap $\rightarrow$ $c = 5$ or 10.

Results: *w/o using data*, next slide..

Example 3: RPPA

Study: Korenblau et al. (2006), protein abundance in bone marrow (BMA) vs. peripheral blood (PB), using RPPA

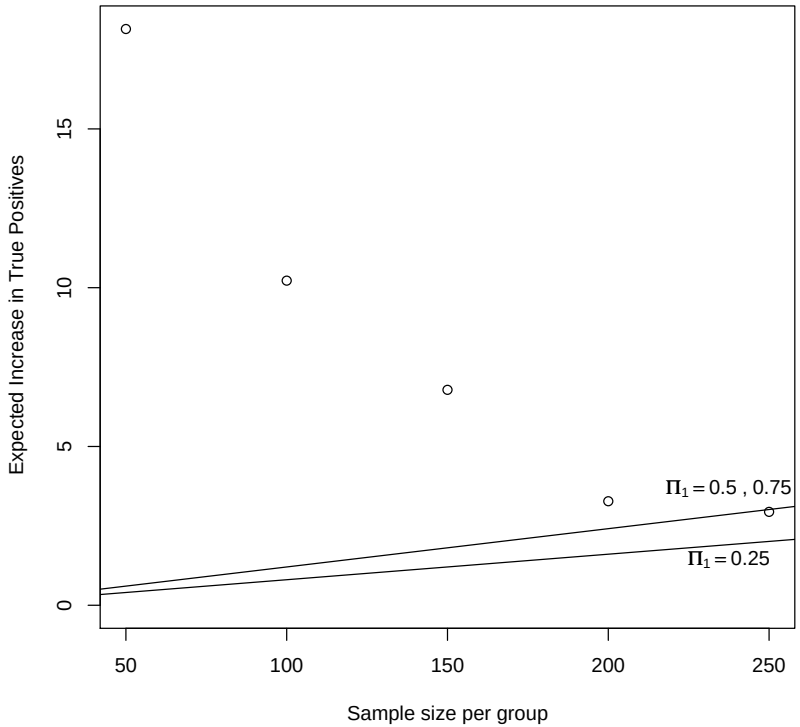
Data: 168 proteins, 435 BMA + 282 PB samples.

Design: batches of 50 samples per group, max of $T = 250$. RPPA is cheap $\rightarrow$ $c = 5$ or 10.

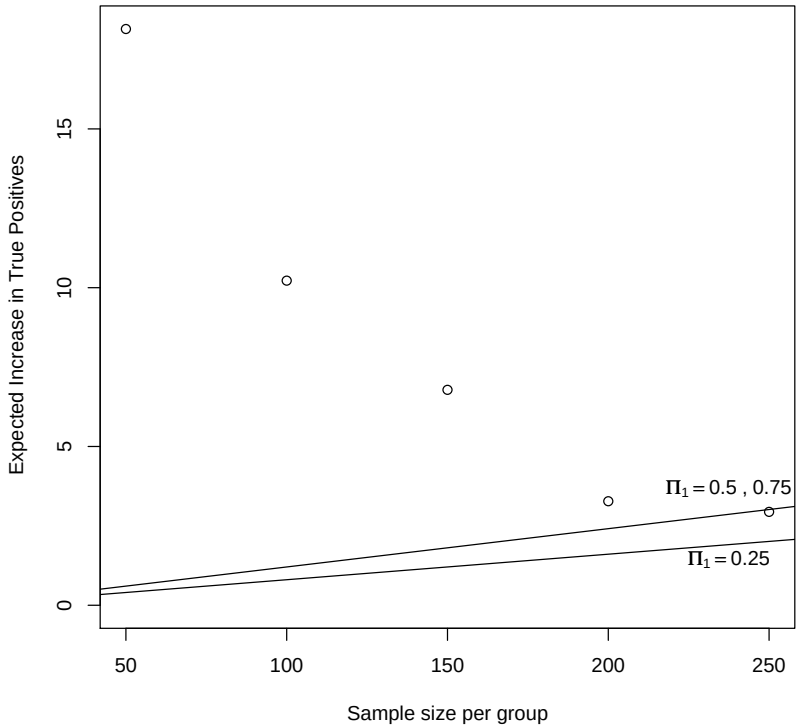
Results: *w/o using data*, next slide..

$x_{ij} = \exp(superCurve)$, Super-Curve pre-processed data (Shannon et al. 2009, Bioinformatics)

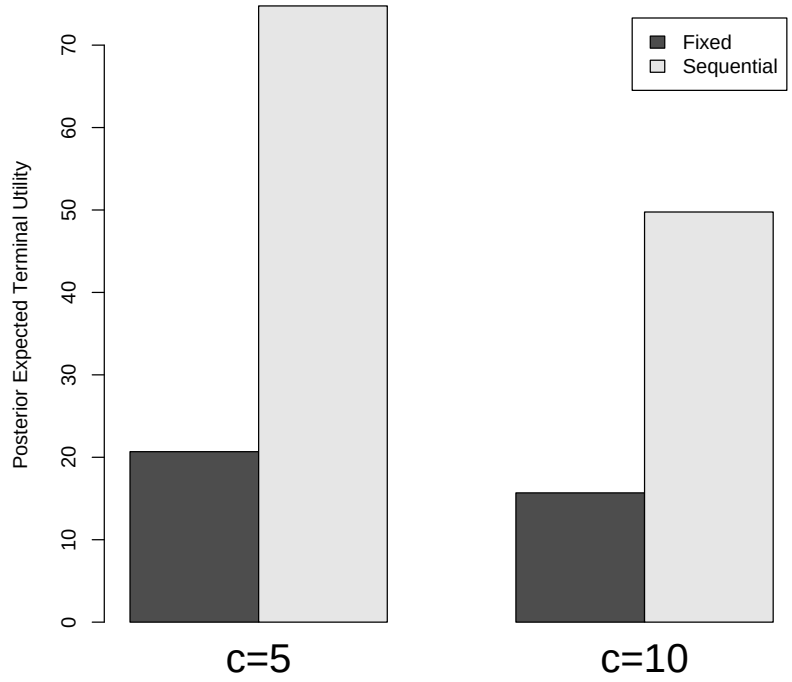
RPPA – Results



RPPA – Results



decision boundary



post expected terminal utility

Summary

- Approx sequential design by decision boundaries
- Need informative priors (post predictive | pilot data; historical data)
- Sequential stopping makes design robust against bad prior choices.
- Utility function, can reflect study goals
- Examples were group comparison, but could be anything: classification, clustering, network discovery ...