Approximate Bayesian Computation for Cancer Simulator

Work Presentation

Probabilistic Machine Learning Group, Aalto University School of Science

Divyat Mahajan, Henri Pesonen, Miika Nikula, Samuel Kaski

Introduction

Problem Statement

- Extend the work on ABC for Cancer Simulator by inferring two additional parameters: Prob. of Birth and Prob. of Death
- Personalized cancer treatment represented as a complex box function with no analytic expression for likelihood available
- Simulating data for particular values of parameters is possible but the likelihood expression is intractable
- Likelihood Free Inference methods for estimating the posterior of parameters
- Task it to infer the patient specific parameters using 3 weeks of data and use them to predict the outcomes for the end of treatment

Cancer Simulator

Mathematical formulation: $X_\theta = f(\theta)$

- A stochastic simulator generating time series data
- f is the black box function representing the treatment procedure.
- No closed form expression for Likelihood: $p(X|\theta)$
- Possible to draw simulations: $X_\theta \sim p(X|\theta)$
- Cancer Simulator can be interpreted as a black box generative model
- The goal is to infer the parameters using time series data reduced to 3 weeks and make predictions about the data at the end of 12 weeks

Parameters(θ) of the Cancer Simulator

- Cell Cycle Length, ChemoSensitivity, Probability of Birth, Probability of Death

Output(X) of the Cancer Simulator:

- Time Series of the following 5 variables indexed at 30 min intervals: Cancer Cells, Vessels, Vegf, O_2, Avastin
- Dimension of Data: $(336 * T, 5)$ where T: Number of weeks of treatment

Literature Review

Rejection ABC Algorithm

Central Idea:

- Posterior approximated by sampling from the prior over parameters and by selecting those values that lead to low discrepancy between simulated and observed data

Algorithm 1 Rejection ABC algorithm for estimating N independent samples from approximate posterior distribution

1: **for** $i = 1$ to N **do**
2: **while** $d(y_\theta, y_o) <= \epsilon$ **do**
3: $\theta \sim$ Prior p(θ)
4: $y_\theta \sim$ from Cancer Simulator
5: **end while**
6: $\theta^i \leftarrow \theta$
7: **end for**

Approximate Posterior: $p_{d,\epsilon}(\theta|y_0) \propto p(d(y_\theta, y_o) \leq \epsilon)p(\theta)$

Issues with ABC

- Due to high dimensionality of data, often informative summary statistics that reduce dimensionality need to be designed to compute distances i.e. $d(T(y_\theta), T(y_0) \leq \epsilon)$
- Quality of inference depends a lot on the summary statistics
- Sufficient statistics cannot always be used due to curse of dimensionality for ABC. Constructing a set of statistics that are informative enough is important
- Rejection ABC algorithm suffers from the drawback of huge rejection rate, lack of knowledge about which parameter space region is close to the true values
- High computational cost of ABC since most parameter values are likely to result in large distances
- The cost of a simulating data may also be very high in some cases

Issues with ABC

- MCMC ABC and SMC ABC construct proposal distributions to improve on the issue of sampling directly from prior distribution
- Sampling from the proposal distribution $q(x|x^i)$ and accepting samples with probability $A(x|x^i)$, where $x^i = (\theta^i, y^i)$ is the ith state of Markov chain such that $d(y^i, y_0) \leq \epsilon$
- Choosing $q(x|x^i) = q(\theta|\theta^i)p(y|\theta)$, it can be shown that: $A(x|x^i) = A(\theta|\theta^i)I(d(y, y_o) \leq \epsilon)$
- The acceptance criterion is a coupling of the MH sampling criteria for θ and ABC rejection criterion
- It still suffers from the high rejection rate due to ABC rejection criterion: $d(y, y_o) \leq \epsilon$

Bayesian Optimisation for ABC

- Learn an approximate model for likelihood by modelling the relation between parameters and the discrepancy between observed and simulated data
- A non parametric approximation of the likelihood by a kernel density estimate:

$$L_K(\theta) = E(K(\phi_o, \phi_\theta)) \tag{1}$$

$$L_\kappa(\theta) = E(\kappa(\Delta_\theta)) \tag{2}$$

$$L_\kappa(\theta) \geq \kappa(J(\theta)) \tag{3}$$

  where the κ is a convex non-negative function, $J(\theta) = E(\Delta_\theta)$, $\kappa(\Delta_\theta) = K(\phi_o - \phi_\theta)$
- Estimate the expectation in $J(\theta)$ by solving the regression problem with variable as θ and response variable as Δ_θ
- A suitable choice for κ as uniform density kernel leads to an approximation of likelihood:

$$L(\theta) = cP(\Delta_\theta < h) \tag{4}$$

  where c is scaling parameter and h is bandwidth of kernel κ

Bayesian Optimisation for ABC

- Solve the regression problem $\Delta_\theta = f(\theta)$ by actively constructing the training set $\{\theta_i, \Delta_{\theta_i}\}$
- Model f by a Gaussian Process Prior $\mathcal{N}(m_t(\theta), K_t(\theta))$ and with evidence set at time t being $\epsilon_t = \{(\theta_1, \Delta_{\theta_1}), ..., (\theta_t, \Delta_{\theta_t})\}$, the posterior for f is available in closed form $\mathcal{N}(\mu_t(\theta), \nu_t(\theta))$
- Select new point θ_{t+1} by minimising the Acquisition Function $A_t(\theta) = \mu_t(\theta) - \eta_t * \sqrt{\nu_t(\theta)}$
  (η_t gives the trade off between exploration and exploitation)
- With the learnt gaussian process, $J(\theta) = E(\Delta_\theta) = \mu_t(\theta)$
- Approximation for likelihood: $L(\theta) = c * F(\frac{h - \mu_t(\theta)}{\sqrt{\nu_t(\theta)}})$

Bayesian Optimisation for ABC

- This approach can specifically focus on regions in the parameter space where the discrepancy Δ_θ tends to be small by minimizing the acquisition function. This reduces the number of runs of the simulator required to estimate posterior.
- This approach also allows to incorporate smoothness assumption: Points nearby in parameter space should have similar discrepancy with observed data
- With analytic expression for likelihood, its easy to draw independent samples from posterior using MCMC. This saves a lot of computational cost, since in previous approaches you need to have more runs of the simulator to have more samples from the posterior

Exploratory Analysis

Designing Summary Statistics

- The data output from the cancer simulator is high dimensional (1009 length vector for 3 weeks duration)
- For effective similarity comparison, the dimensionality of data needs to be reduced by evaluating informative statistics
- Design Statistics by simulating data for different combination of parameters and observing the change in data curves in response to change in parameter values
- Experiment design:
  - Data used for summary statistics reduction and inference consists only of Total Cancer Cells from the output from the Cancer Simulator
  - Data simulated over a 6*6*10 size grid of parameters (TCell, ChemoSensitivity, PDeath) with the range as TCell: (1,21), ChemoSensitivity: (1,4), PBirth: (0,1), PDeath: 0.001
  - Simulated data is mean of 5 runs of cancer simulator for each combination of parameters

Issues with simulating data

- It is not possible to simulate data for all combinations of (PBirth, PDeath) due to convergence errors in newton solver. For high values of PDeath, its not possible to simulate data for all possible values of PBirth in the range(0,1)
- The error is caused by a bug in the Cancer Simulator code due to the lack of exception handling when Vessel Cells become zero
- The above issue was taken into consideration while designing the experiments. The experiments ahead focus on simulating data with fixed PDeath=0.001 and a range of other parameters (TCell, ChemoSen, PBirth)

Summary Statistics

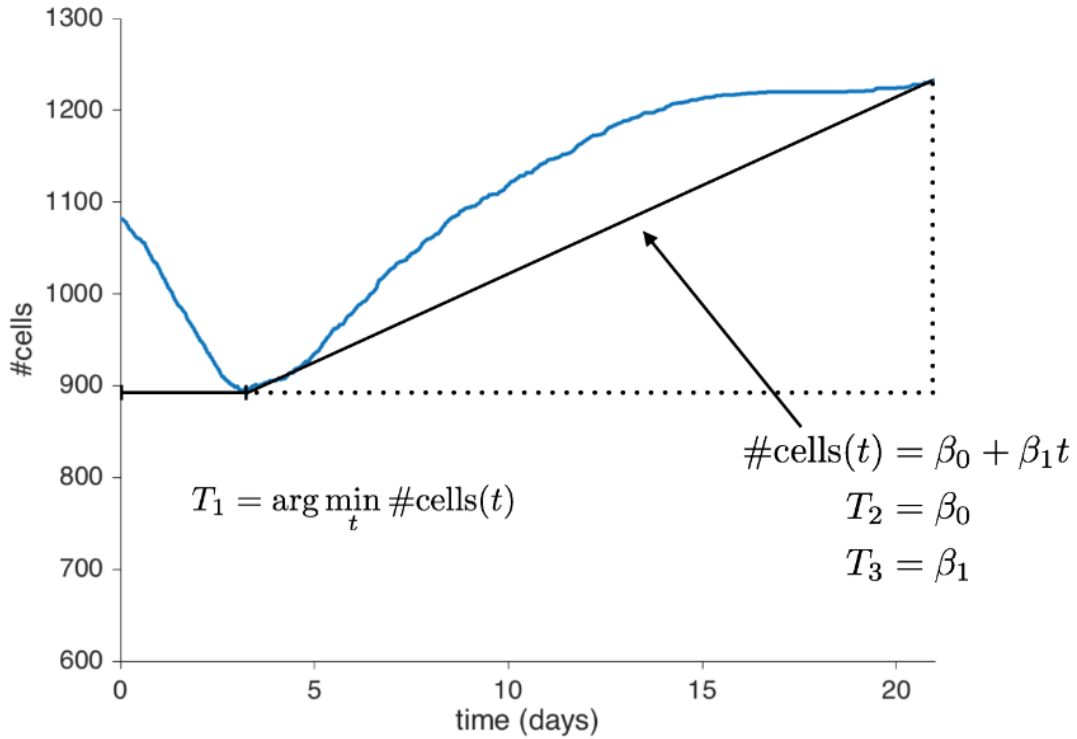
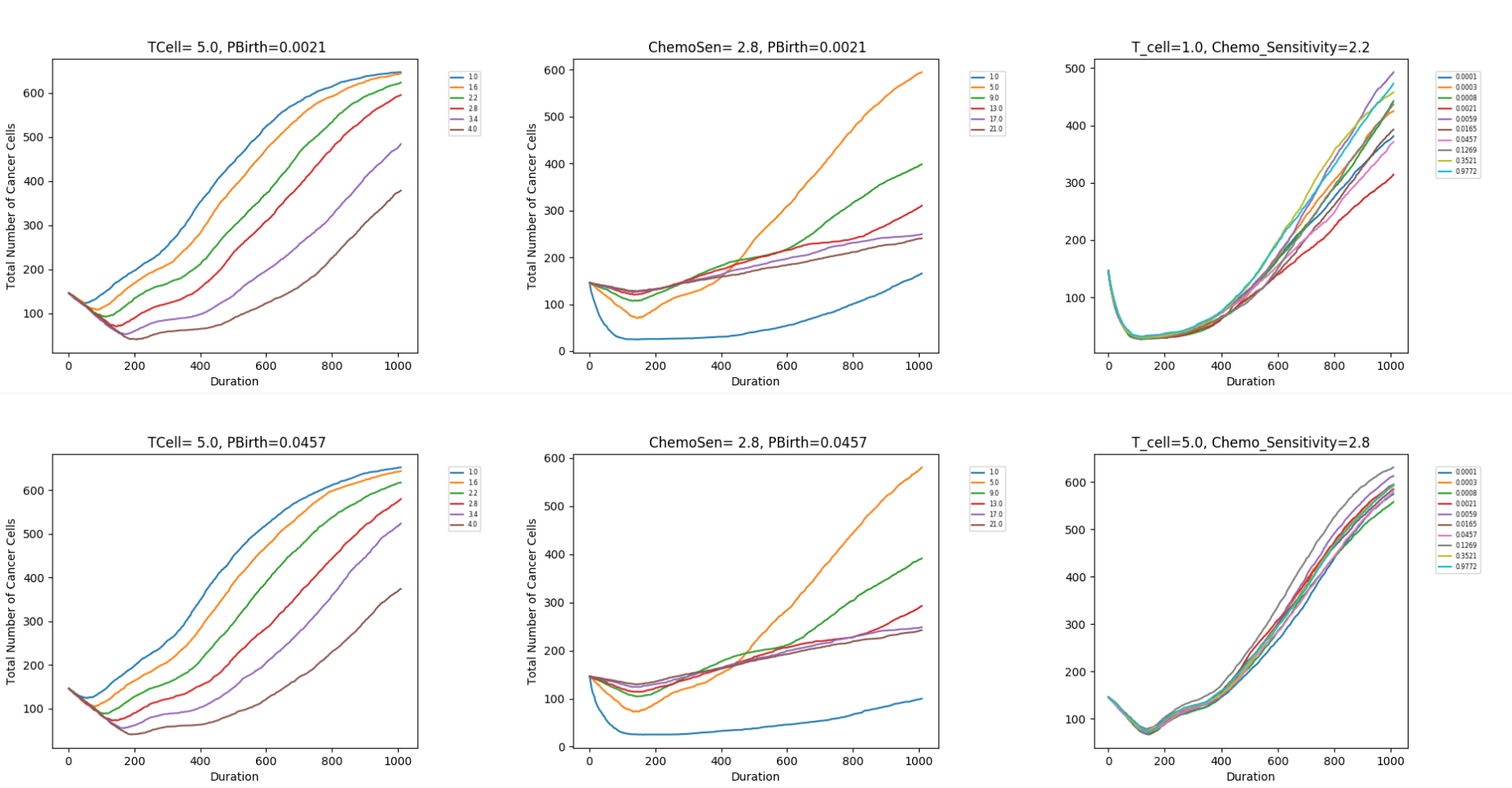


Figure 3: The choice of a set of summary statistics.

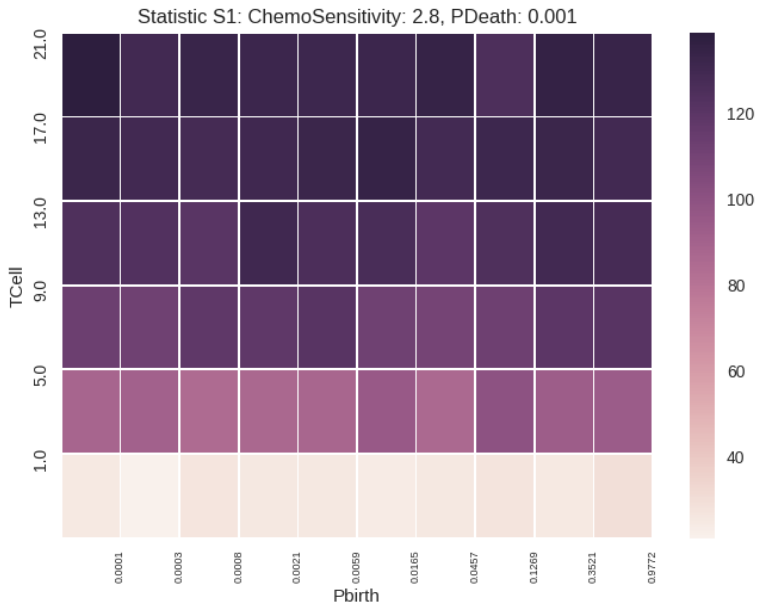
- A slight change of notation in the presentation ahead. The set of Statistics (T1,T2,T3) above would be named (S1,S2,SC3) ahead.

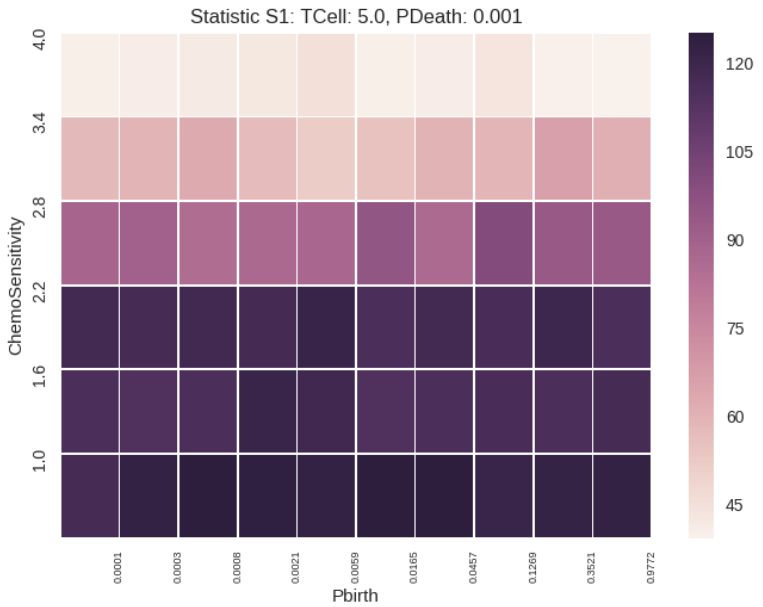
Cancer Cells Data Plot

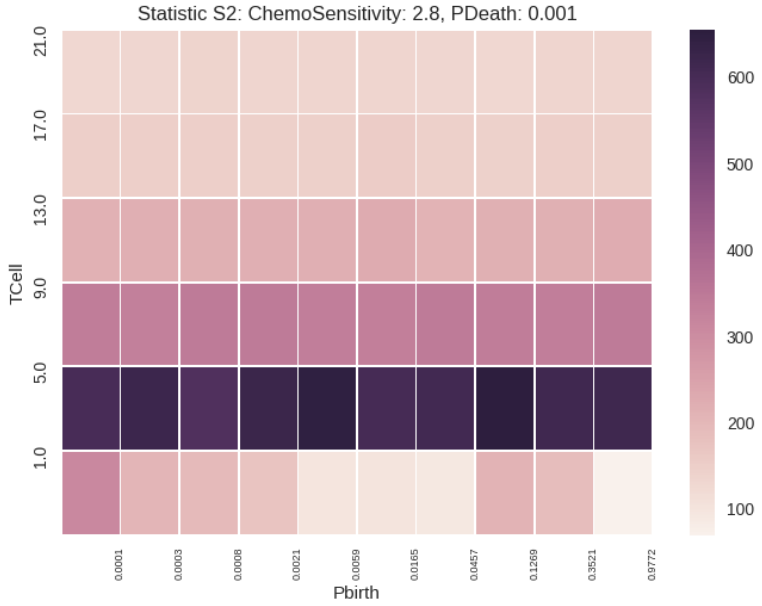


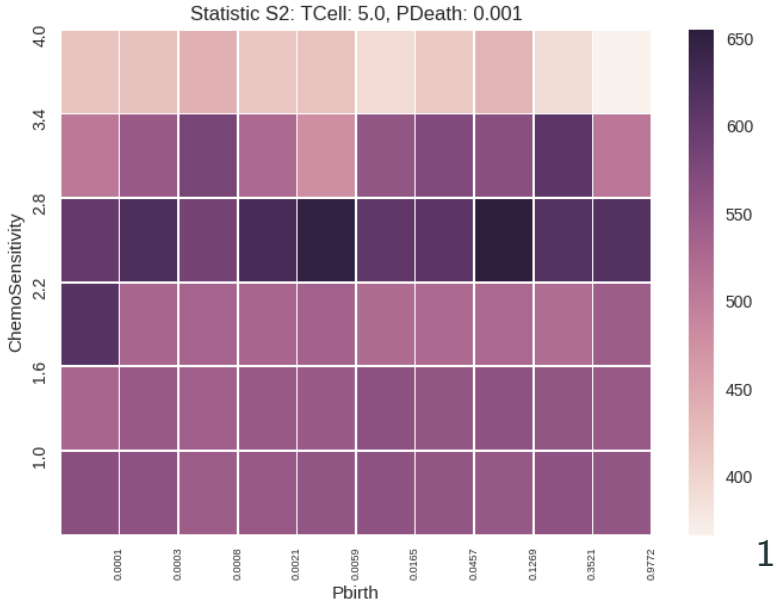
- Point of the minima of cancer cells increases with increasing ChemoSensitivity
- Minimum number of cancer cells increases with increasing TCell
- Less variation in data curves in response to change in Pbirth as compared to changes in ChemoSen, TCell

Statistic Evaluation

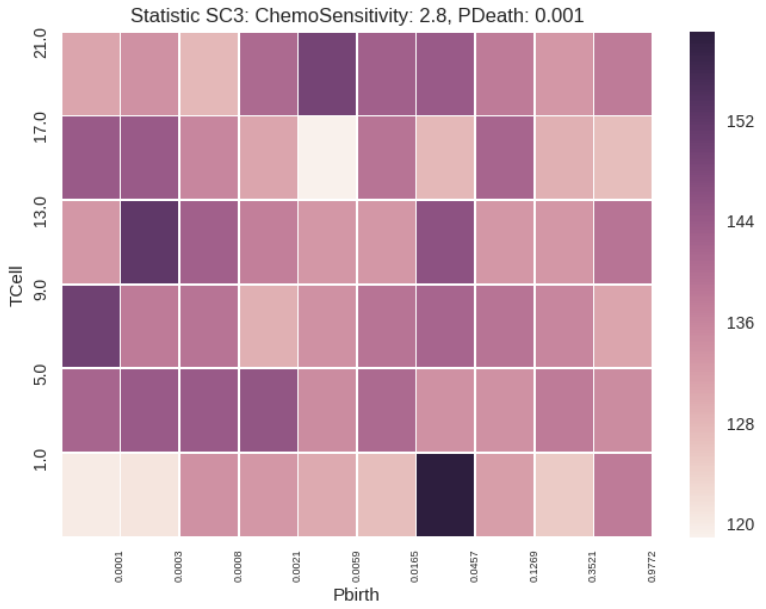


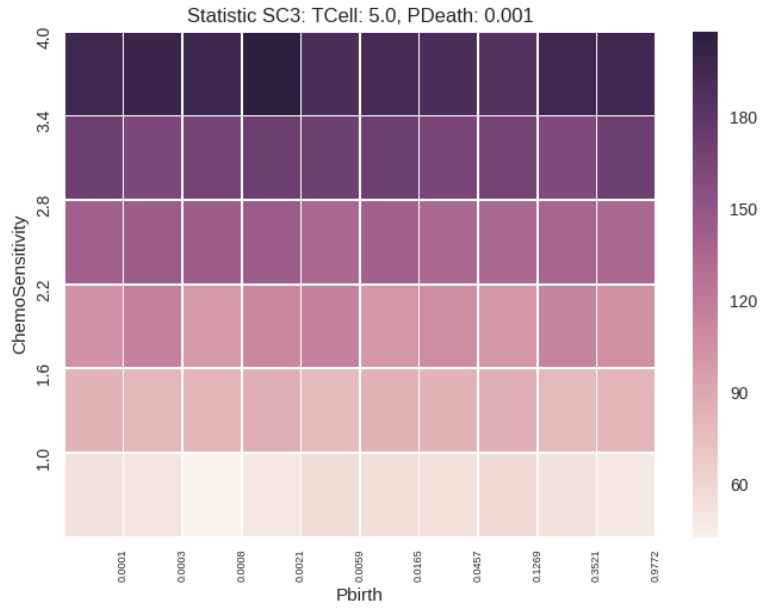


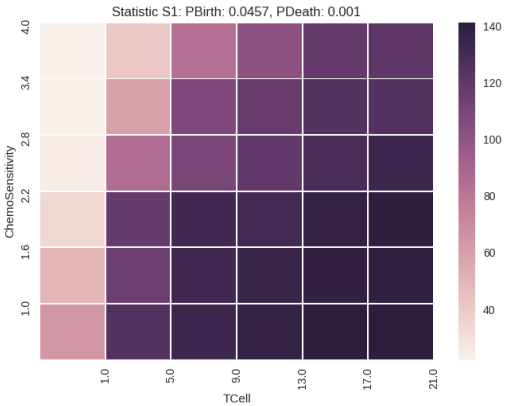


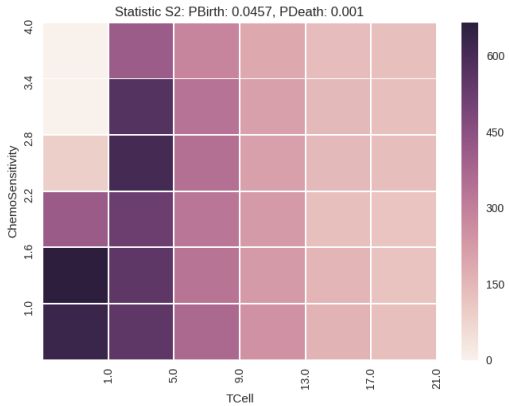


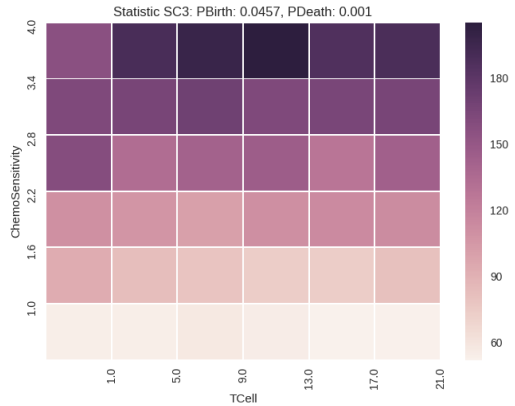
Statistic Evaluation



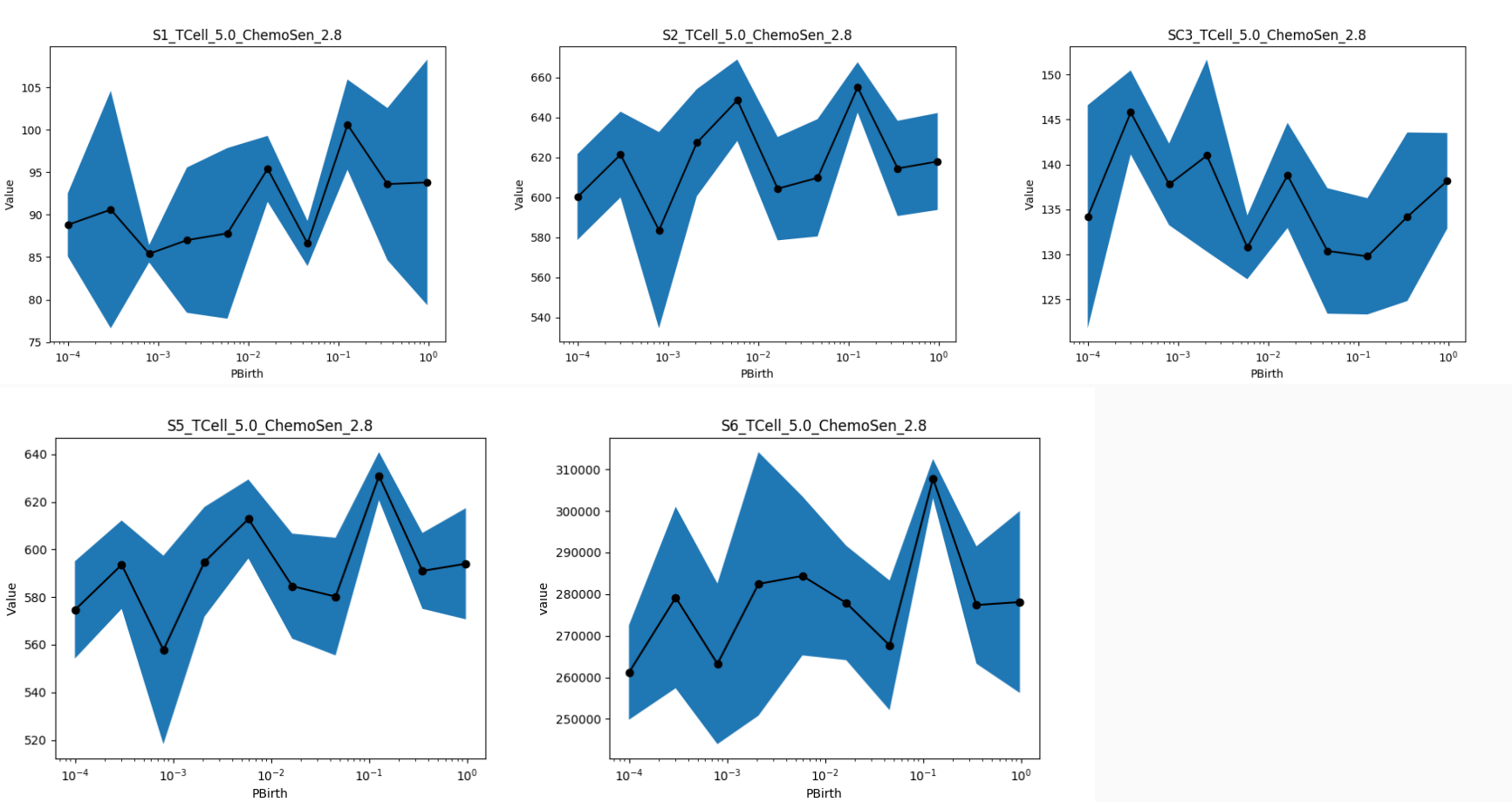








Statistic Evaluation

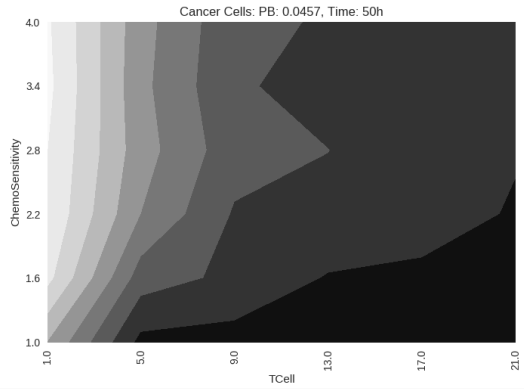


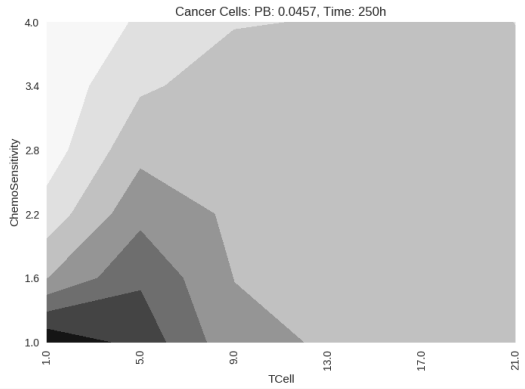
- S5 and S6 are new summary statistics designed with the aim to capture variation with change in PBirth
- S5: Cancer Cells at the end of 3 weeks
- S6: Sum of Cancer Cells during the whole duration of 3 weeks

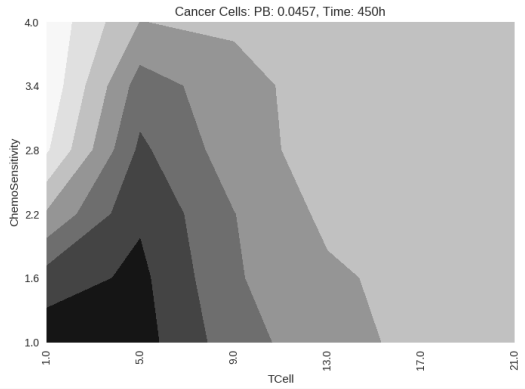
Observations

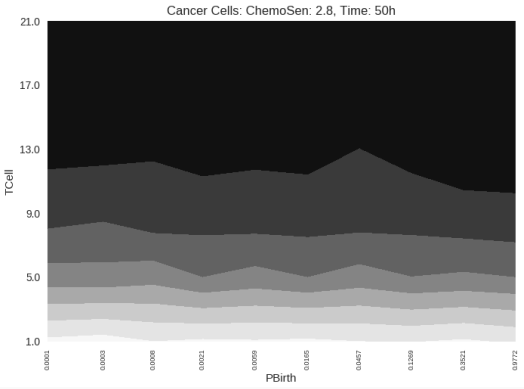
- Current Statistics capture the trend for change in response to ChemoSensitivity parameter in the region with lower values of TCell
- Current Statistics do not effectively capture variation in response to change in the parameter PBirth
- The change in Statistics values in response to change in PBirth seems quite random and does not represent any relation between them
- There does not seem to be any derivable property that changes monotonically with response to change in PBirth with other parameters fixed
- Cancer Cell Evolution plots also show that cancer cells(**without any reduction via summary statistics**) show random variations with changes to PBirth parameter

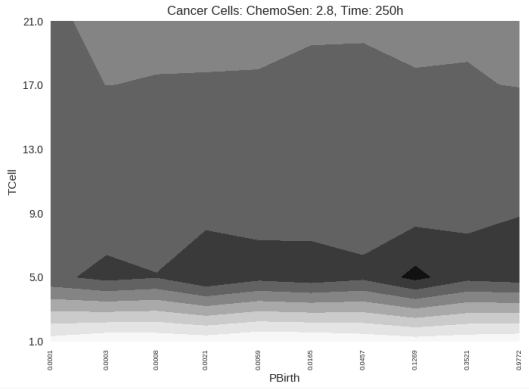
Cancer Cell Evolution

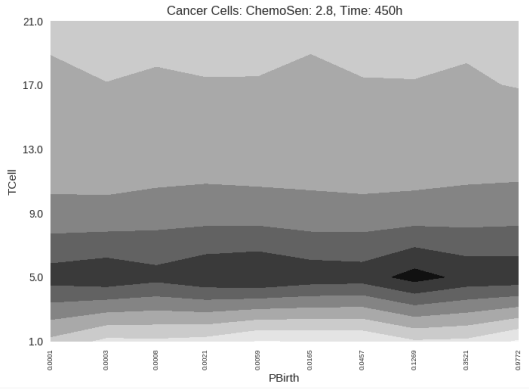


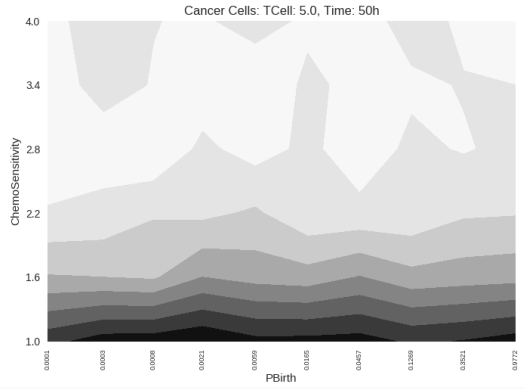


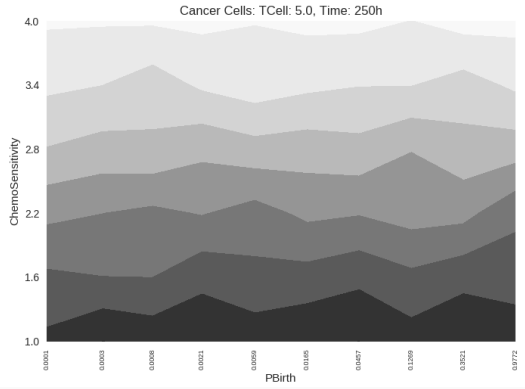


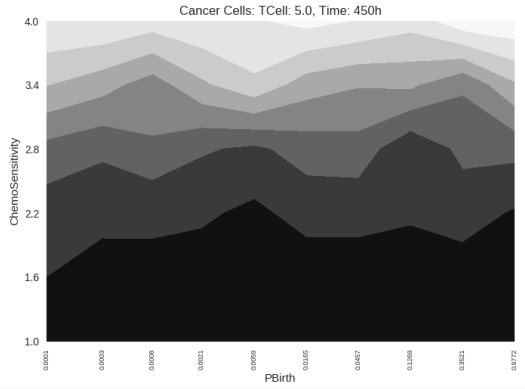












Inference using BOLFI

Experiment Design

- Using ELFI(Engine for Likelihood Free Inference) to implement Bayesian Optimization framework for the inference of parameters
- Plots ahead estimate posterior over parameters by using MCMC to draw 2000 samples(rejecting the first 1000 samples as burn in period of MCMC)
- The prior over parameters is defined as a uniform distribution with range for each parameter as follows:
  - TCell: Uniform(1,21)
  - ChemoSen: Uniform(1,4)
  - PBirth: Uniform(0,1)
- Discrepancy between observed and simulated data(modelled by a Gaussian Process) for the purpose of constructing training set is calculated using Euclidean Distance between Summary Statistic of observed and simulated data

$$\Delta_\theta = ||d(\phi_0, \phi_\theta)||$$

- $\phi_\theta = (S1(y_\theta), S2(y_\theta), S3(y_\theta))$, where S1,S2,S3 are the summary statistics as defined previously

Results

- Inferring (TCell, ChemoSen) with their true values as (6.26, 3.0) and fixed PBirth=0.002, PDeath=0.001
  Mean of samples from BOLFI: (6.17, 3.51)
- Inferring PBirth with fixed other parameters (TCell, ChemoSen, PDeath) as (6.26, 3.0, 0.001):

True Value	Mean Estimate
0.004	0.51
0.01	0.49
0.1	0.47

- Inferring (ChemoSen, TCell, PBirth) with fixed Death as 0.001:

True Value	Mean Estimate
(1.6, 5.0, 0.01)	(2.0, 7.0, 0.50)
(1.6, 5.0, 0.04)	(2.16, 9.31, 0.48)
(1.6, 5.0, 0.1)	**(2.1, 6.2, 0.44)**
(1.6, 5.0, 0.35)	(1.97, 2.34, 0.41)
(1.6, 5.0, 0.97)	(1.90, 17.8, 0.48)

Results

- Inferring (ChemoSen, TCell, PBirth) with fixed Death as 0.001:

True Value	Mean Estimate
(2.8, 1.0, 0.0002)	(3.1, 13.8, 0.56)
(2.8, 1.0, 0.0008)	**(3.8, 1.4, 0.46)**
(2.8, 1.0, 0.002)	(3.6, 8.2, 0.40)
(2.8, 1.0, 0.005)	(3.1, 12.3, 0.47)
(2.8, 1.0, 0.01)	(3.4, 16.7, 0.48)
(2.8, 1.0, 0.04)	(3.7, 14.2, 0.45)
(2.8, 1.0, 0.13)	**(3.8, 1.2, 0.60)**
(2.8, 1.0, 0.35)	(3.5, 7.19, 0.47)
(2.8, 1.0, 0.97)	**(3.9, 1.3, 0.62)**

Comments

- Inferring (Tcell, ChemoSen) results in good estimates, consistent with the previous study done by Henri
- Inferring PBirth alone or inferring jointly (TCell, ChemoSen, PBirth) leads to very poor estimates for PBirth
- Joint Inference of (TCell, ChemoSen, PBirth) leads to an interesting point, **the quality of estimate of TCell depends strongly on the true value of PBirth**. The relation between them seems to be quite random though

Conclusion

Possible Outcomes

Following outcomes for additional parameters (Prob Birth, Prob Death) are possible from the results of experiments performed:

- The parameters cannot be inferred using current summary statistics
- The parameters cannot be inferred using 3 weeks of data.
- The parameters do not have a significant systematic effect on Cancer Treatment
- The parameter PBirth cannot be inferred for the specific case of PDeath=0.001
- The parameters do not have systematic effect in case of Non Responsive patients

Future Work

- Designing new statistics that capture the variation in a better way with response to change in the Pbirth parameter
- Inferring parameters using different data modality: Vessels, O2, Avastin or Vegf in place of Cancer Cells
- Exploratory analysis for the full grid containing a complete range of PDeath values
- Simulating the same experiment for the case of a Responsive Patient