An application of data assimilation to mathematical biology

Eric J. Kostelich

kostelich@asu.edu

ARIZONA STATE UNIVERSITY

SCHOOL OF **MATHEMATICAL & STATISTICAL SCIENCES**

SIAM Applied Dynamical Systems Conference

Snowbird, UT

May 21, 2017

Collaborators

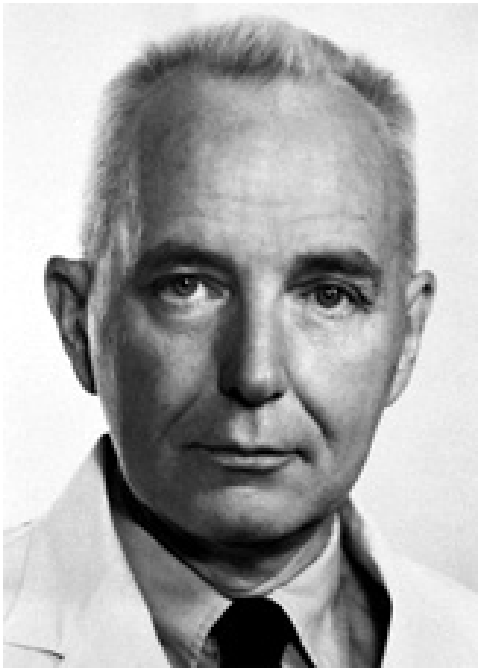
- Yang Kuang, Javier Baez, ASU
- Alan Bryce, MD, Mayo Clinic Arizona

Overview of prostate cancer

- American men have $\sim 1/7$ lifetime risk ($\sim 233{,}000$ new cases and $\sim 30{,}000$ deaths per year)
- At diagnosis, it is difficult to distinguish aggressive tumors from indolent ones
- Overtreatment is probably common—and risky
- Prostate-specific antigen (PSA) is an imprecise proxy for cancer
- Stage C: Spread to seminal vesicles
- Stage D: Spread to lymph nodes and/or bones

Charles Huggins (1901–1997)

- Discovered in the ’30s and ’40s that castration causes regression of prostate cancer tumors—thus proving that some cancers are hormone dependent



Charles Huggins, 2

- Hormone dependence was later demonstrated for most breast cancers
- Huggins shared the 1966 Nobel Prize in Medicine for this discovery

Therapy of prostate cancer

- Castration (flutamide) remains the standard of care for locally metastatic tumors—but side effects can be debilitating
- Androgen deprivation therapy usually is very effective initally, but tumors eventually evolve resistance
- Intermittent therapy attempts to minimize side effects and forestall resistance
- The first clinical trial (Akakura et al., 1992) involved 7 patients whose serum PSA and androgen levels were measured clinically approximately once per month

Portz, Nagy, and Kuang model (2012)

- This model postulates the existence of "androgen dependent" (AD) and "androgen independent" (AI) cell populations
- Basic idea: Growth under resource constraints (Droop, 1968)
- Prostate cancer cells require a minimum level $q_{\min}$ of androgen for survival
- Suppose $Q_1(t)$ is the cell quota for androgen at time t

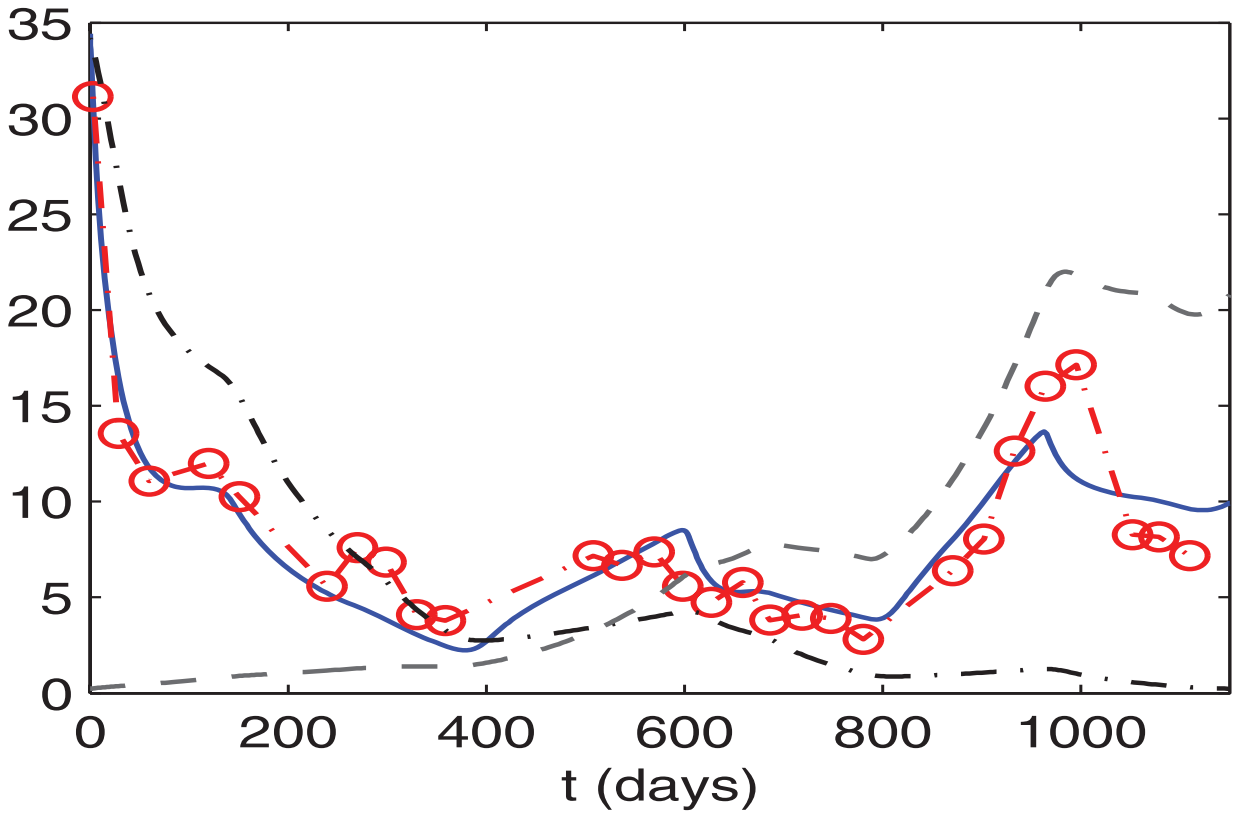
Portz, Nagy, and Kuang model (2012), 2

- Growth model for androgen-dependent (AD) cells:

$$\frac{dX_1}{dt} = \mu\left(1 - \frac{q^1_{\min}}{Q_1(t)}\right) X_1(t) + \text{mutations}$$

- Examples: If $Q_1(t) = 2q^1_{\min}$, then $X_1(t) \sim e^{\mu t/2}$
- $Q_1(t) = q^1_{\min}/2 \Longrightarrow X_1(t) \sim e^{-\mu t}$
- $Q_1(t) = q^1_{\min}/10 \Longrightarrow X_1(t) \sim e^{-9\mu t}$
- Of course, $q^i_{\min}$ isn't known and probably varies by patient
- The model includes equations for $X'_{1,2}(t)$ and $Q'_{1,2}(t)$ with 19 total parameters

Sample result: Case 4 in Portz *et al.* (2012)



Prediction problem: Will another cycle of therapy be effective?

Comments on the fitting procedure

- MATLAB's fminsearch finds a set of parameters and initial conditions that fits the observations reasonably well
- Thus one can argue that the model is clinically validated

Comments on the fitting procedure

- MATLAB's fminsearch finds a set of parameters and initial conditions that fits the observations reasonably well
- Thus one can argue that the model is clinically validated
- Prediction: Another round of therapy is likely to fail

Comments on the fitting procedure

- MATLAB's fminsearch finds a set of parameters and initial conditions that fits the observations reasonably well
- Thus one can argue that the model is clinically validated
- Prediction: Another round of therapy is likely to fail
- Criticism: The procedure is *post hoc* and does not ascribe a level of confidence in the results
- Data assimilation may provide updated estimates of treatment efficacy during the course of therapy

Five necessary ingredients for data assimilation

1. A dynamical forecast model
2. An estimate of the forecast uncertainties
3. Observations of the process under study
4. Estimates of the observational errors
5. A forward operator ($\mathbf{H}$: model space $\rightarrow$ observations)

These ingredients yield a local ensemble transform Kalman filter (LETKF)

Ingredient #1: The Portz et al. (2012) model

- Androgen-dependent (AD) cells:

$$\frac{dX_1}{dt} = \mu\left(1 - \frac{q^1_{\min}}{Q_1(t)}\right) X_1(t) + \text{mutations}$$

- Androgen-independent (AI) cells:

$$\frac{dX_2}{dt} = \mu\left(1 - \frac{q^2_{\min}}{Q_2(t)}\right) X_2(t) + \text{mutations}$$

- The cell quota depends on the availability of androgen:

$$\frac{dQ_i}{dt} = \beta_i(A) - \gamma_i(A)Q_i, \quad i = 1, 2$$

where β and γ depend on the serum androgen level A

Ingredient #1: The Portz et al. (2012) model, 2

- PSA production:

$$\frac{dP}{dt} = \sigma_1(Q_1)X_1 + \sigma_2(Q_2)X_2 - \delta P$$

- Portz model state vector: $\mathbf{x} = (X_1, X_2, Q_1, Q_2, P)$

Ingredient #2: Estimates of forecast uncertainty

- Start with k guesses of the initial condition
- Integrate the model forward to the first forecast time (1 month in our case)
- At each model output time, we have a corresponding set of "background" state vectors

$$\left\{\mathbf{x}_b^i\right\}_{i=1}^k \quad \text{with ensemble mean} \quad \bar{\mathbf{x}}_b$$

- Each state vector is ℓ-dimensional ($\ell = 5$ for this model)

Ingredient #2: Estimates of forecast uncertainty, 2

- Form the $\ell \times k$ matrix $\mathbf{X}_b$ of ensemble perturbations whose ith column is $\mathbf{x}_b^i - \bar{\mathbf{x}}_b$
- Let $\mathbf{w} \in \mathbb{R}^k$. Then

$$\mathbf{x} = \bar{\mathbf{x}}_b + \mathbf{X}_b \mathbf{w}$$

  corresponds to a model state vector in the span of the ensemble
- For instance, $\mathbf{w} = \mathbf{e}_i$ yields $\mathbf{x} = \mathbf{x}_b^i$

Ingredient #2: Estimates of forecast uncertainty, 2

- Form the $\ell \times k$ matrix $\mathbf{X}_b$ of ensemble perturbations whose ith column is $\mathbf{x}_b^i - \bar{\mathbf{x}}_b$
- Let $\mathbf{w} \in \mathbb{R}^k$. Then

$$\mathbf{x} = \bar{\mathbf{x}}_b + \mathbf{X}_b \mathbf{w}$$

  corresponds to a model state vector in the span of the ensemble
- For instance, $\mathbf{w} = \mathbf{e}_i$ yields $\mathbf{x} = \mathbf{x}_b^i$
- Lemma: If $\mathbf{w}$ is Gaussian with mean $\mathbf{0}$ and covariance $(k-1)^{-1}\mathbf{I}$, then $\mathbf{x} = \bar{\mathbf{x}}_b + \mathbf{X}_b \mathbf{w}$ is Gaussian with mean $\bar{\mathbf{x}}_b$ and covariance $\mathbf{P}_b = (k-1)^{-1}\mathbf{X}_b \mathbf{X}_b^{\mathrm{T}}$

Ingredient #3: Treatment of observations

- The observations here are PSA levels (ng/mL)
- The Portz model explicitly predicts the serum PSA level
- $\mathbf{H}$ for this model a single component of the model state vector $\mathbf{x}$
- In general, $\mathbf{H}(\mathbf{x})$ is an s-vector of predicted observations given the model state $\mathbf{x}$
- Remember that $\mathbf{x} = \bar{\mathbf{x}}_b + \mathbf{X}_b \mathbf{w}$

Ingredient #3: Treatment of observations, 2

- Each ensemble solution yields a predicted set of observations

$$\left\{\mathbf{y}_b^i = \mathbf{H}(\mathbf{x}_b^i)\right\}_{i=1}^k \quad \text{with mean} \quad \bar{\mathbf{y}}_b$$

- Define the $s \times k$ matrix $\mathbf{Y}_b$ of observation perturbations whose ith column is $\mathbf{y}_b^i - \bar{\mathbf{y}}_b$
- In this way, we can express obervations as a function of $\mathbf{w}$ by linearization:

$$\mathbf{H}(\mathbf{x}) = \mathbf{H}(\bar{\mathbf{x}}_b + \mathbf{X}_b\mathbf{w}) \approx \bar{\mathbf{y}}_b + \mathbf{Y}_b\mathbf{w}$$

Ingredient #4: Estimates of observational error

- Assumption #1: The variance in the measurements is 1 ng/mL
- Assumption #2: The measurement errors in successive clinic visits are uncorrelated
- The measurement covariance matrix **R** is diagonal

Remarks on ingredient #5, the forward operator

- **H** is the fifth component of the model state vector **x**

The minimization procedure

- Minimize the objective function

$$\begin{aligned}\widehat{J}(\mathbf{w}) &= (k-1)^{-1}\mathbf{w}^{\mathrm{T}}\mathbf{w} + \\ &\quad [\mathbf{y} - \bar{\mathbf{y}}_b - \mathbf{Y}_b\mathbf{w}]^{\mathrm{T}}\mathbf{R}^{-1}[\mathbf{y} - \bar{\mathbf{y}}_b - \mathbf{Y}_b\mathbf{w}]\end{aligned}$$

The minimization procedure

- Minimize the objective function

$$\begin{aligned}\widehat{J}(\mathbf{w}) &= (k-1)^{-1}\mathbf{w}^{\mathrm{T}}\mathbf{w} + \\ &\quad [\mathbf{y} - \bar{\mathbf{y}}_b - \mathbf{Y}_b\mathbf{w}]^{\mathrm{T}}\mathbf{R}^{-1}[\mathbf{y} - \bar{\mathbf{y}}_b - \mathbf{Y}_b\mathbf{w}]\end{aligned}$$

- The minimizer is the k-vector $\mathbf{w}_a = \mathbf{Q}\mathbf{Y}_b^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y} - \bar{\mathbf{y}}_b)$ where $\mathbf{Q} = [(k-1)\mathbf{I} + \mathbf{Y}_b^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{Y}_b]^{-1}$

The ensemble update

- First update the ensemble mean:

$$\bar{\mathbf{x}}_a = \bar{\mathbf{x}}_b + \mathbf{X}_b \mathbf{w}_a$$

- Then update the anslysis ensemble:

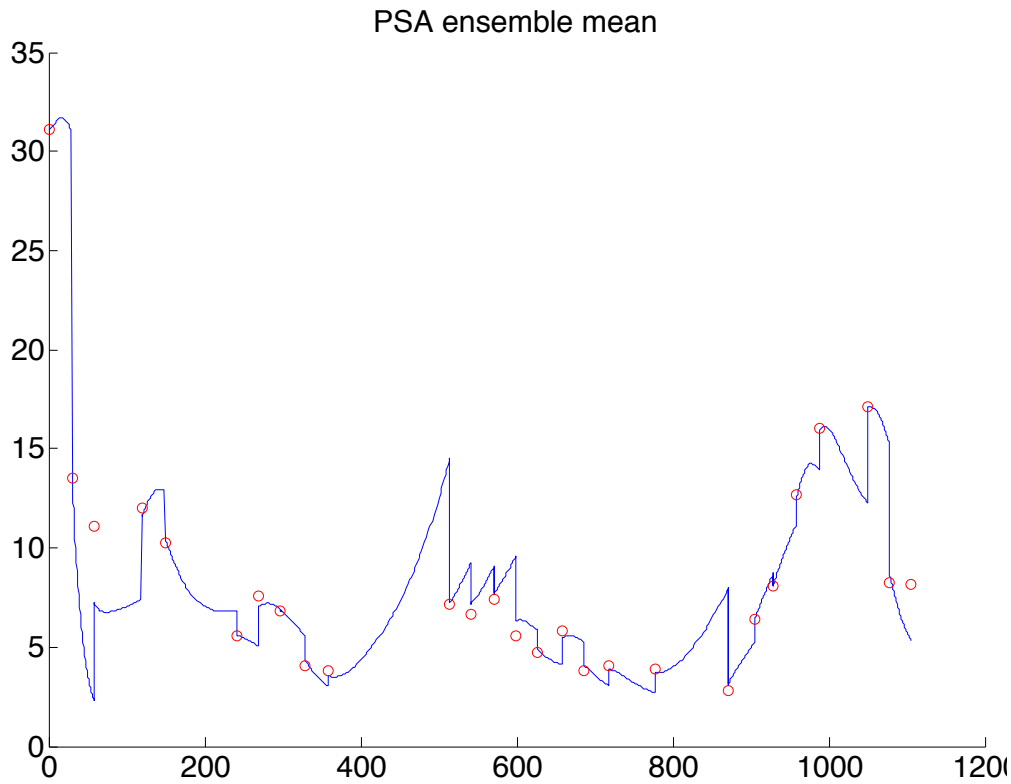
$$\mathbf{X}_a = \mathbf{X}_b \mathbf{W}_a$$

  where $\mathbf{W}_a = [(k-1)\mathbf{Q}]^{1/2}$ (symmetric square root)
- We obtain $\mathbf{x}_a^i$ by adding $\bar{\mathbf{x}}_a$ to the ith column of $\mathbf{X}_a$
- The analysis covariance matrix is $\mathbf{P}_a = \mathbf{X}_b \mathbf{Q} \mathbf{X}_b^{\mathrm{T}}$

Sketch of the algorithm

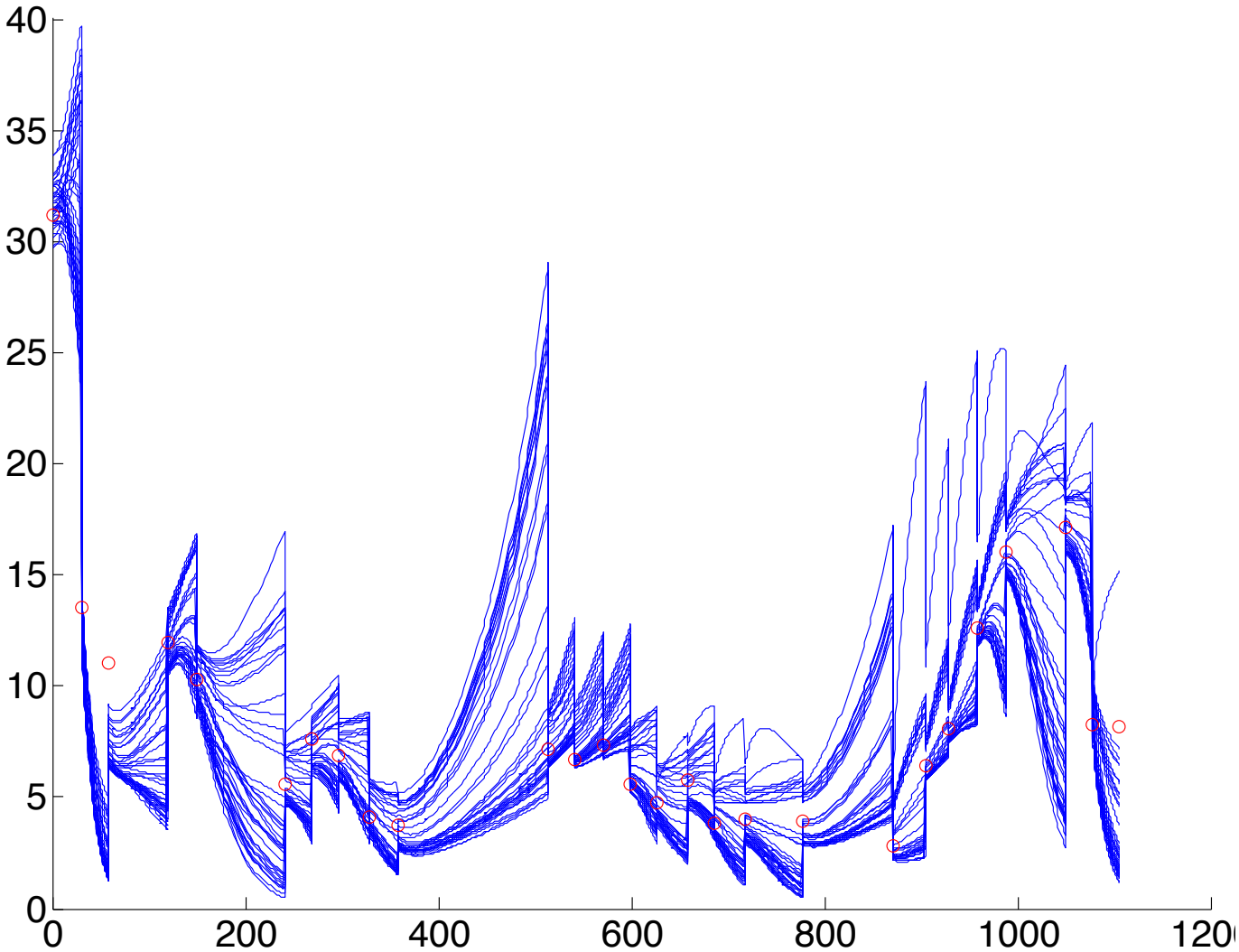
- Step 1: Make an ensemble of forecasts ("background") from a set of reasonable initial conditions from $t = t_n$ to $t = t_{n+1}$, the next observation time
- Step 2: At t_{n+1}, collect a set of observations (PSA levels)
- Step 3: Update the forecast ensemble with the Kalman estimate
- Step 4: $n \leftarrow n + 1$ and go back to Step 1
- Typically (but not always!) $t_{n+1} - t_n = 1$ month

Sample result—Serum PSA for Case 4

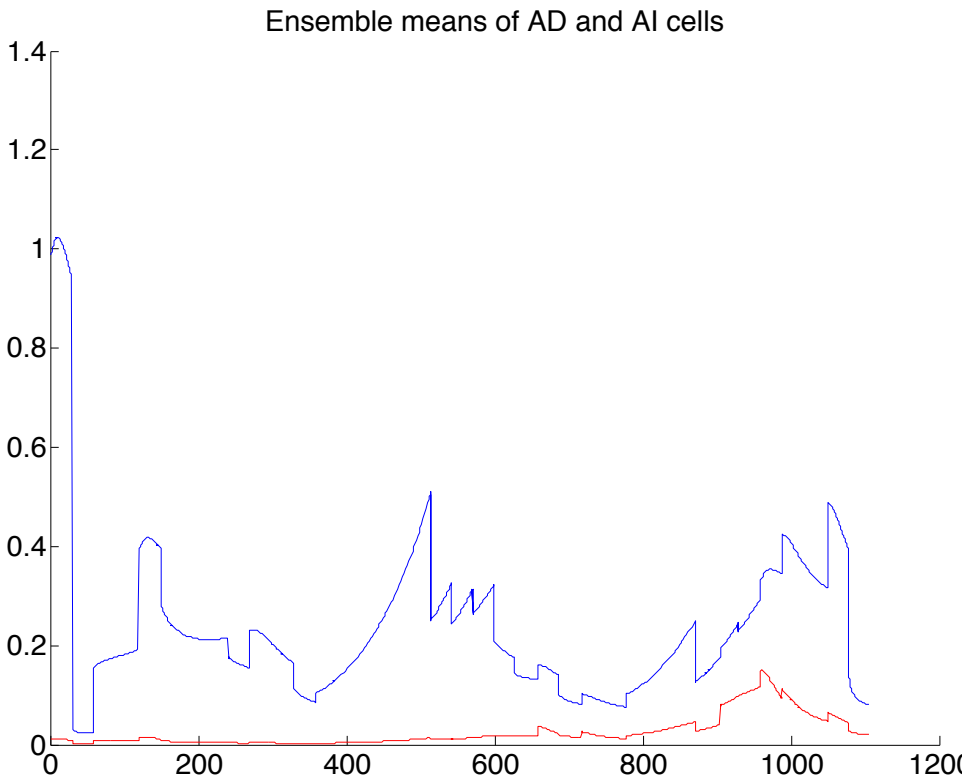


- Blue curve: ensemble mean of model PSA levels
- Red circles: clinical measurements

Case 4 PSA ensemble



Case 4 cell population means



- Blue: AD cells; Red: AI cells
- Prediction: Another round of therapy will work

Why do we get two different predictions?

- The state vector involves two tumor subpopulations and associated cell quotas, plus PSA
- The only component that is directly observable is PSA
- PSA production: $P' = \sigma_1(Q_1)X_1 + \sigma_2(Q_2)X_2 - \delta P$
- The cell quotas Q_i depend on serum androgen A, which is roughly constant except immediately before and after suppression therapy
- Net result: PSA levels do not depend uniquely on X_1 and X_2
- Thus, the tumor subpopulations are unidentifiable from PSA levels alone!

Conclusions

- The Local Ensemble Transform Kalman Filter is a flexible tool that accounts for empirical model and observational uncertainty
- One's ability to estimate the state of a process depends on the dynamical model and on the measurements
- Caveat: If the ensemble covariance does not accurately reflect the actual model uncertainties, then the filter can diverge

Further caveats

- George Box: "Essentially, all models are wrong, but some are useful"
- Models are most useful when they are designed with the observables in mind
- Open problems include how best to handle model errors
- Prostate tumor cells may dedifferentiate
- As the disease progresses, patients may have large tumors that produce very little PSA

Some references

- Mathematical details: B. R. Hunt, E. K., I. Szunyogh, *Physica D* **230** (2007) 112–126
- Portz model: T. Portz, Y. Kuang and J. D. Nagy, *AIP Advances* **2** (2012), 011002