Spatial Sampling for Model Selection

Nabil Hajj Chehade, Greg Pottie
EE, UCLA

Introduction

Environment Modeling Uncertainty

- Wireless sensing systems are very useful for applications where we need to learn about environmental phenomena over spatial and temporal fields.
- Parametric models are widely used to represent these environmental fields and to answer questions and inferences regarding the phenomena.
- Choosing a model structure to represent the field involves a great deal of uncertainty.
- Often data are collected at random locations. We present methods for data collection to optimize the desired inferences.
- Often a single model M is used. If M does not characterize a phenomenon correctly, the inferences and predictions will not be accurate.
- It is better to start with multiple plausible models and select the model by collecting measurements at informative locations.

Regression

- The polynomial regression model:
  \[ t = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n + e \]
- Vector form:
  \[ t = X \hat{\beta} + e \]
- We assume Gaussian noise, the ML estimate of \( \theta \) is given by
  \[ \hat{\theta} = (X^T X)^{-1} X^T t \]
- The estimation error covariance matrix is given by
  \[ \text{ Cov}(\hat{\theta}) = (X^T X)^{-1} \sigma^2 \]
- The error covariance matrix depends on the design matrix X and does not depend on the measurements.

Reducing Uncertainty in Model Estimation

Problem Description: Optimal Sensor placement
Where should we collect measurements to optimally estimate the parameters of the regression model?

Assumptions: The model used is the correct model.
- Gaussian noise (\( C = I \)).

Idea: Find the locations that result in the “smallest” error covariance matrix.

Technically:

\[ \min_{t^T I = 1} \det(X \text{diag}(t)) \]

\[ 1 \geq 0 \]

Algorithm: D-Designs
The minimization problem is convex and can be solved with any convex optimization software.

Reducing Uncertainty in Model Selection

Problem Description: Optimal Sensor placement
Where should we collect measurements to optimally choose a model that represents the field?

Assumptions: A set of plausible models.
- The set contains the correct model.
- Gaussian noise.

Idea: Find the locations where the “difference” between the two models is the largest.

Derivation:

Two model case

\[ M_1: t = \eta_1(x, \theta_1) + e_1, \quad i = 1, \ldots, n \]

\[ M_2: t = \eta_2(x, \theta_2) + e_2, \quad i = 1, \ldots, n \]

This is a binary hypothesis test and the probability of error is given by

\[ \max_{\hat{\epsilon}} \left\| \text{diag}(\hat{\epsilon}) \left( (x, \hat{\epsilon}, e) \right) \left( (x, \hat{\epsilon}, e) \right)^T \right\| \]

where \( \hat{\epsilon} = \arg \min \left\| \text{diag}(\hat{\epsilon}) \left( (x, \hat{\epsilon}, e) \right)^T \right\| \)

\[ \hat{\epsilon} = \arg \min \left\| \text{diag}(\hat{\epsilon}) \left( (x, \hat{\epsilon}, e) \right)^T \right\| \]

Algorithm: T-Designs [1]

1. Two model case:

   1. Given a design \( X \), where \( N \) is the number of observations, find
      \[ \theta_{\text{true}} = \arg \min_{\theta} \sum_{i=1}^{N} \left( \eta_i(x_i, \theta) - y_i \right)^2 \]
      \[ \hat{\theta}_{\text{true}} = \arg \min_{\theta} \sum_{i=1}^{N} \left( \eta_i(x_i, \hat{\theta}) - y_i \right)^2 \]
   2. Add to the design a point \( x_{\text{new}} \) such that:
      \[ x_{\text{new}} = \arg \max_{x} \left( \eta(x, \hat{\theta}) - \eta(x, \theta_{\text{true}}) \right) \]
   3. The \( (N+1) \) observation is taken at \( x_{\text{new}} \).
   4. Go back to 1.

2. Multiple model case:

   1. Given a design \( X \), where \( N \) is the number of observations, for each model \( \eta_i(x, \theta) \) find:
      \[ \hat{\theta}_{i,j} = \arg \min_{\theta} \sum_{i=1}^{N} \left( \eta_i(x_i, \theta) - y_i \right)^2 \]
   2. Rank the models by goodness of fit, for example
      \[ L_1, L_2, \ldots, L_r \]
   3. Add to the design a point \( x_{\text{new}} \) such that:
      \[ x_{\text{new}} = \arg \max_{x} \left( \eta(x, \hat{\theta}_{i,j}) - \eta(x, \hat{\theta}_{i,j}) \right) \]
   4. The \( (N+1) \) observation is taken at \( x_{\text{new}} \).
   5. Update \( Z_{\text{new}} = \left[ \left( \frac{1}{N+1} \right)^* \hat{\theta}_{i,j} + \frac{N}{N+1} \hat{\theta}_{i,j} \right] \)
   6. Go back to 1.

Evaluation:

\[ M_1: t = \theta_1 x + \theta_{21} x_1 + \theta_{22} x_2 + \theta_{23} x_3 + \theta_{24} x_4 + e_1, \quad i = 1, \ldots, n \]

\[ M_2: t = \theta_2 x + \theta_{21} x_1 + \theta_{22} x_2 + \theta_{23} x_3 + \theta_{24} x_4 + e_2, \quad i = 1, \ldots, n \]

Likelihood:

\[ M_2: M_2 = 2.18910 \]

Future Work

Model Estimation:
- Optimal designs for robust model estimation.
- Optimal designs for multiple modality fields while incorporating the correlation between the modalities.
- Spatio-temporal fields.

Model Selection:
- Investigate the effect of the prior probabilities on the design.
- Extend the work presented to situations when the set of models considered does not include the correct model.