CPSC 540: Machine Learning

Hierarchal Bayes

Mark Schmidt

University of British Columbia

Winter 2020

Hierarchical Bayesian Models

- Type II maximum likelihood is not really Bayesian:
  - We're dealing with w using the rules of probability.
  - But we're treating λ as a parameter, not a nuissance variable.
    - You could overfit λ.

- Hierarchical Bayesian models introduce a hyper-prior $p(\lambda \mid \gamma)$.
  - We can be "very Bayesian" and treat the hyper-parameter as a nuissance parameter.

- Now use Bayesian inference for dealing with λ:
  - Work with posterior over λ, $p(\lambda \mid X, y, \gamma)$, if integral over w is easy.
  - Or work with posterior over w and λ.
  - You could also consider a Bayes factor for comparing λ values:

$$p(\lambda_1 \mid X, y, \gamma)/p(\lambda_2 \mid X, y, \gamma),$$

  which now account for belief in different hyper-parameter settings.

Model Selection and Averaging: Hyper-Parameters as Variables

- Bayesian model selection ("type II MAP"): maximizes hyper-parameter posterior,

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}}\, p(\lambda \mid X, y, \gamma)$$
$$= \underset{\lambda}{\operatorname{argmax}}\, p(y \mid X, \lambda) p(\lambda \mid \gamma),$$

  further taking us away from overfitting (thus allowing more complex models).
  - We could do the same thing to choose order of polynomial basis, σ in RBFs, etc.
- Bayesian model averaging considers posterior predictive over hyper-parameters,

$$\hat{y}^i = \underset{\hat{y}}{\operatorname{argmax}} \int_{\lambda} \int_{w} p(\hat{y} \mid \hat{x}^i, w) p(w, \lambda \mid X, y, \gamma) dw d\lambda.$$

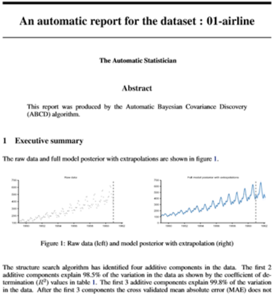
- Could maximize marginal likelihood of hyper-hyper-parameter γ, ("type III ML"),

$$\hat{\gamma} = \underset{\gamma}{\operatorname{argmax}}\, p(y \mid X, \gamma) = \underset{\gamma}{\operatorname{argmax}} \int_{\lambda} \int_{w} p(y \mid X, w) p(w \mid \lambda) p(\lambda \mid \gamma) dw d\lambda.$$

Application: Automated Statistician

- Hierarchical Bayes approach to regression:
  1. Put a hyper-prior over possible hyper-parameters.
  2. Use type II MAP to optimize hyper-parameters of your regression model.

- Can be viewed as an automatic statistician:
  `http://www.automaticstatistician.com/examples`



An automatic report for the dataset : 01-airline

The Automatic Statistician

Abstract

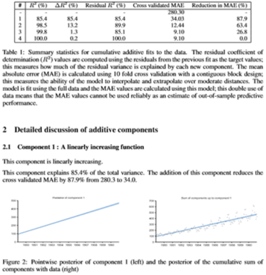
This report was produced by the Automatic Bayesian Covariance Discovery (ABCD) algorithm.

1 Executive summary

The raw data and full model posterior with extrapolations are shown in figure 1.

Figure 1: Raw data (left) and model posterior with extrapolation (right)

The structure search algorithm has identified four additive components in the data. The first 2 additive components explain 98.5% of the variation in the data as shown by the coefficient of determination (R^2) values in table 1. The first 3 additive components explain 99.8% of the variation in the data. After the first 3 components the cross validated mean absolute error (MAE) does not



#	R^2 (%)	ΔR^2 (%)	Residual R^2 (%)	Cross validated MAE	Reduction in MAE (%)
-	-	-	-	280.30	-
1	85.4	85.4	85.4	34.03	87.9
2	98.5	13.2	89.9	12.44	63.4
3	99.8	1.3	85.1	9.10	26.8
4	100.0	0.2	100.0	9.10	0.0

Table 1: Summary statistics for cumulative additive fits to the data. The residual coefficient of determination (R^2) values are computed using the residuals from the previous fit as the target values; this measures how much of the residual variance is explained by each new component. The mean absolute error (MAE) is calculated using 10 fold cross validation with a contiguous block design; this measures the ability of the model to interpolate and extrapolate over moderate distances. The model is fit using the full data and the MAE values are calculated using this model; this double use of data means that the MAE values cannot be used reliably as an estimate of out-of-sample predictive performance.

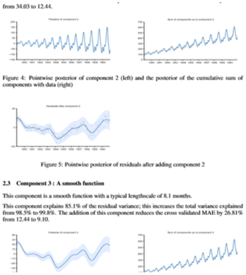
2 Detailed discussion of additive components

2.1 Component 1 : A linearly increasing function

This component is linearly increasing.

This component explains 85.4% of the total variance. The addition of this component reduces the cross validated MAE by 87.9% from 280.3 to 34.0.

Figure 2: Pointwise posterior of component 1 (left) and the posterior of the cumulative sum of components with data (right)



from 34.03 to 12.44.

Figure 4: Pointwise posterior of component 2 (left) and the posterior of the cumulative sum of components with data (right)

Figure 5: Pointwise posterior of residuals after adding component 2

2.3 Component 3 : A smooth function

This component is a smooth function with a typical lengthscale of 8.1 months.

This component explains 85.1% of the residual variance; this increases the total variance explained from 98.5% to 99.8%. The addition of this component reduces the cross validated MAE by 26.81% from 12.44 to 9.10.

Discussion of Hierarchical Bayes

- "Super Bayesian" approach:
  - Go up the hierarchy until model includes all assumptions about the world.
  - Some people try to do this, and have argued that this may be how humans reason.
- Key advantage:
  - Mathematically simple to know what to do as you go up the hierarchy:
    - Same math for w, z, λ, γ, and so on (all are nuissance parameters).
- Key disadvantages:
  - It can be hard to exactly encode your prior beliefs.
  - The integrals get ugly very quickly.

Hierarchical Bayes as a Graphical Model

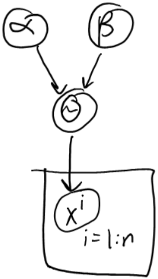
- Let x^i be a binary variable, representing if treatment works on patient i,

$$x^i \sim \text{Ber}(\theta).$$

- As before, let's assume that θ comes from a beta distribution,

$$\theta \sim \mathcal{B}(\alpha, \beta).$$

- We can visualize this as a graphical model:

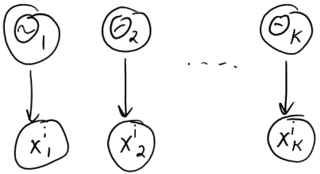


Hierarchical Bayes for Non-IID Data

- Now let x^i represent if treatment works on patient i in hospital j.
- Let's assume that treatment depends on hospital,

$$x_j^i \sim \text{Ber}(\theta_j).$$

- So the x_j^i are only IID given the hospital.



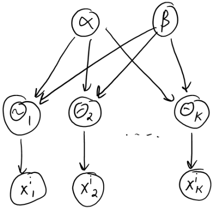
- Problem: we may not have a lot of data for each hospital.
  - Can we use data from one hospital to learn about others?
  - Can we say anything about a hospital with no data?

Hierarchical Bayes for Non-IID Data

- Common approach: assume the θ_j are drawn from common prior,

$$\theta_j \sim \mathcal{B}(\alpha, \beta).$$

- This introduces dependency between parameters at different hospitals:

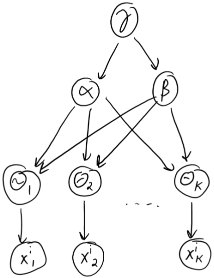


- But, if you fix α and β then you can't learn across hospitals:
  - The θ_j and d-separated given α and β.

- Type II MLE would optimize α and β given non-IID data.

Hierarchical Bayes for Non-IID Data

- Consider treating α and β as random variables and using a hyperprior:



- Now there is a dependency between the different θ_j (for unknown α and β).

- Now you can combine the non-IID data across different hospitals.
  - Data-rich hospitals inform posterior for data-poor hospitals.
  - You even consider the posterior for new hospitals with no data.

Summary

- Hierarchical Bayes goes even more Bayesian with prior on hyper-parameters.
  - Leads to Bayesian model selection and Bayesian model averaging.
- Relaxing IID assumption with hierarchical Bayes.
- Next time: modeling cancer mutation signatures.