Bayesian Mixture of Inverse Regression (BMI)

Code

BMI.tar.gz is the code.

A Brief Tutorial

Unzip the downloaded file. To start, first make sure that the folder "BMI/" and the sub folders and files are in the search path of Matlab. The input training data should at least consist of a response vector Y, a covariate matrix X with rows observations and columns explanatory/input variables, an integer number specifying the number of e.d.r. directions, and a structure specifying whether the response is continuous or discrete and the corresponding parameters to be tuned. The program will return the posterior mean and posterior draws for the e.d.r. directions. For a further detailed explanation of the inputs and outputs type "help bmi" in the command line.

Examples

Dimension Reduction

We illustrate how dimension reduction can be performed in this section by considering a classification problem for handwritten digits. An illustration of the digit data is shown in Figure 1. Each digit is represented by a 28*28=784 vector that contains the pixel values.

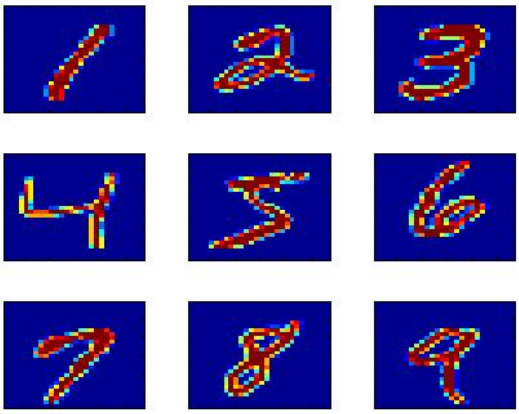

Figure1: Illustration of handwritten digit data 1-9.

Suppose we want to classify digit "5" and digit "8". We can collect 100 samples for each digit and label digit "5" as response=1 and digit "8" as response=0, so that the covariates X is a 200*784 matrix for pixel values and the response Y is a 200 vector for class labels. We can type in the command line:

[B, Bpost]=bmi(Y, X, 1, res, 'T')

where the fourth argument is a structure with elements:

res.type='d': the response is discrete; res.alpha0=1: the concentration parameter = 1

The fifth argument 'T' tells the program to pre-process the data by a principle component analysis, since in this case the number of input variables is larger than the sample size.

Now Bpost, a 784*1000 matrix, contains the posterior samples for the e.d.r., and B is the orthonormalized version of the mean of Bpost. "mean(Bpost,2)" and "std(Bpost,0,2)" return the posterior mean and standard deviation for this direction, both 784 vectors. We can plot them in a visually friendly way by "imagesc(reshape(mean(Bpost,2),28,28)')" and "imagesc(reshape(std(Bpost,0,2),28,28)')" with the results shown in Figure 2. The red part in the left panel is exactly the region that differentiates digit "5" and "8", hence if we project the original data onto this direction we can immediately perform classification. The right panel indicates small uncertainty.

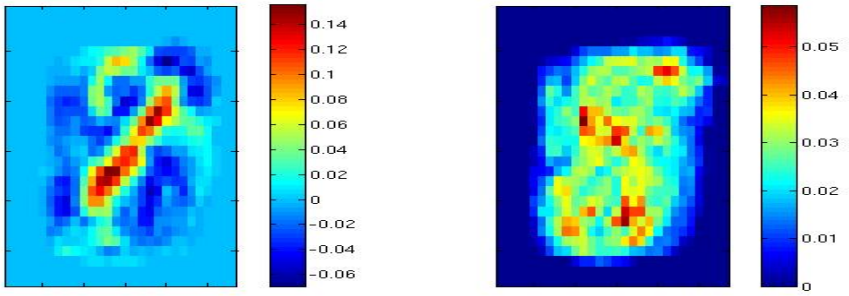

Figure2: The left panel is the posterior mean and the right panel is the posterior standard deviation for the top dimension reduction direction.

For any questions and comments please email km68 at stat dot duke dot edu. Thanks.