Here we quickly discuss three important and related topics in statistics:

* Over-fitting
* Smoothing
* Cross-validation

Rather than using a formal presentation, these topics are discussed in the context of a "Mock *f*-MRI" experiment.

**Mock *f-*MRI**

We analyze simulated *f*-MRI data, meant to reproduce a “mind-reading” experiment.

The subject is either telling the truth (**Condition T**) or lying (**Condition L**) while the *f*-MRI signal is recorded. The task is to predict the condition from the *f*-MRI data, i.e., classify each trial as belonging either to Condition T or to Condition L.

The *f*-MRI data is high-dimensional and very noisy.

The dataset contains 100 trials in condition T, and 100 trials in condition L. Each trial is a 100×100 real-valued array.

Only the first half of the trials of each condition (i.e., 50 trials), called the TRAINING trials, will be used to **construct the model**. The second half of the trials of each condition, called the TEST trials, will be used for **model evaluation**.

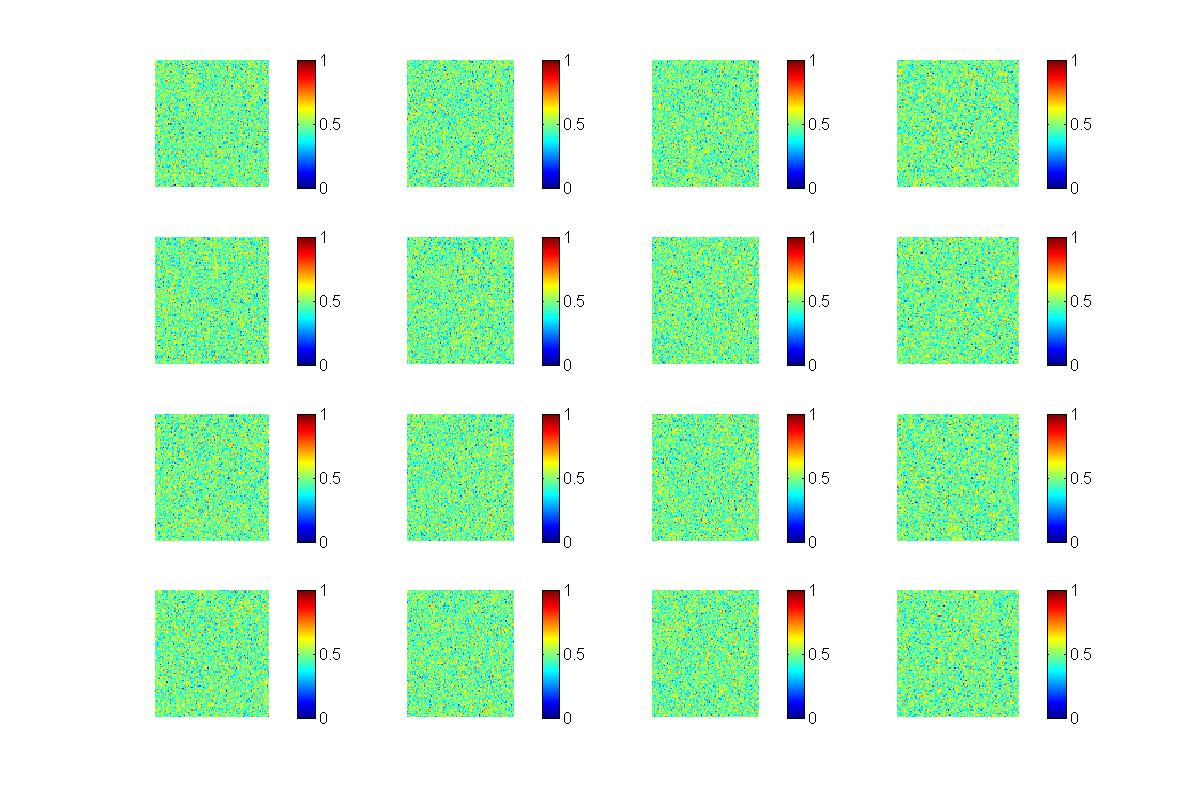
Using two different sets of trials for training and for evaluation is the basis of the statistical method known as *cross-validation*. This method is useful when adjusting the parameters of a *smoothing algorithm*, which is made necessary by the high level of noise in the data.

Given a trial to classify, the classification algorithm computes a *score* from the *f*-MRI data collected during this trial. The score function uses a MODEL, previously constructed from the TRAINING data.

If everything goes well, i.e., if there is enough usable T/L-related signal in the *f*-MRI data, and if a **good model** is used in the scoring function (a good model will turn out to require the correct amount of smoothing), the score tends to be high for L trials, and low for T trials.

The user can then set a *discrimination threshold* at a suitable value *D*. Trials that score less than *D* will be classified T, and trials that score more than *D* will be classified L.

Figure 1 shows 16 of the 100 Condition-T trials (see below). The Condition-L trials are very similar in appearance. There is no clear pattern to be discerned from mere visual examination of the raw data. If there is a pattern, it is buried in the noise. It appears impossible to draw conclusions, from mere visual examination, as to the presence of a signal in the data relevant to the task of predicting the L/T Condition.



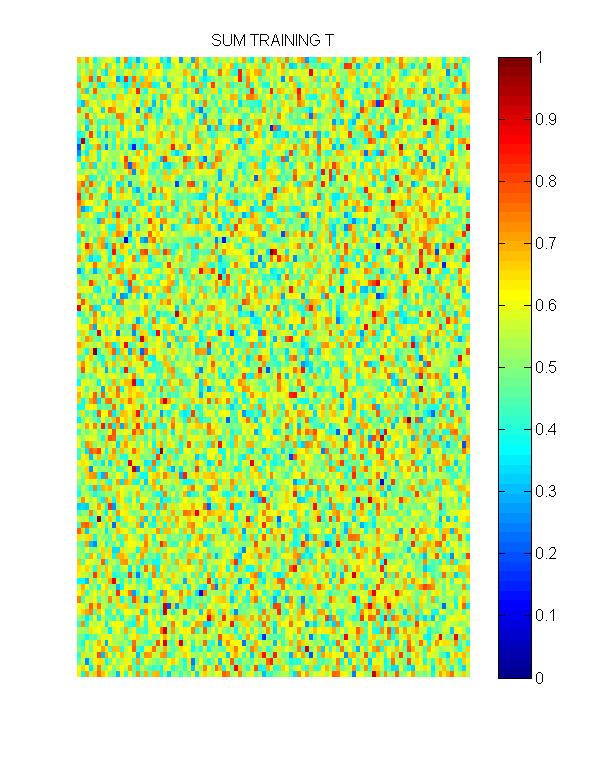
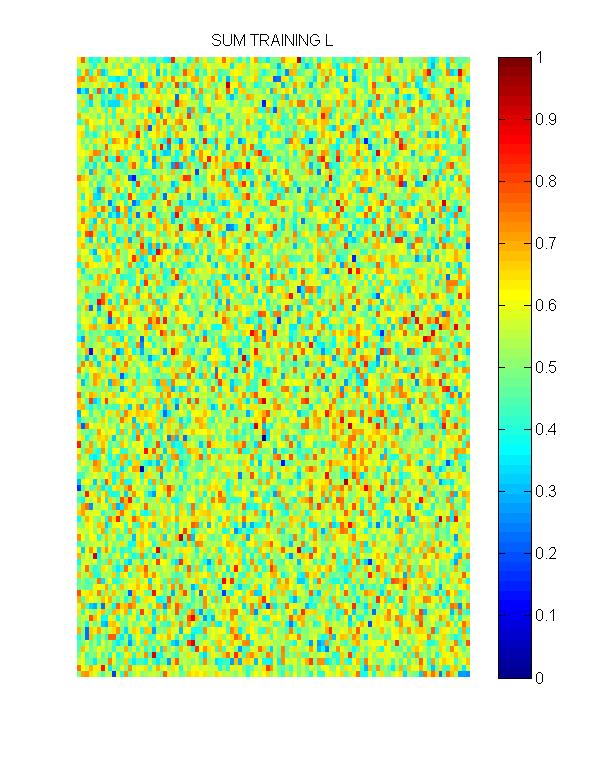
**Figure 1. Condition T single-trial data.**

Figure 2 (below) shows the voxel-by-voxel summation of the first 50 Condition-T trials, i.e., the Training T trials, and Figure 3 shows the voxel-wise sum of the Training L trials.

In these Sum-T and Sum-L images, a few regions seem to be a bit brighter than average. When going back-and-forth between single-trial and summed data and focusing on these regions in the corresponding single-trial images, one may indeed see in some cases a faint trace of the presence of the pattern.

This summing is equivalent to the classical technique of "averaging out the noise," which can be explained as follows. Let us, conceptually, divide the summed data by the number of Training trials, 50. This operation is a mere scaling, hence does not affect visual analysis. We are now considering the *trial-averaged* data in each condition. Averaging leaves the signal unchanged. On the other hand, assuming that, in any given voxel, the noise is independent across trials, averaging divides the amplitude of the noise by √50 ≅ 7. This explains why patterns that are practically invisible in the individual-trial data may become visible in the averaged data.

Since the bright regions in the Sum-L data appear to be different from the bright regions in the Sum-T data, there is some hope that this “signal” could be used for the task of predicting the L/T Condition.

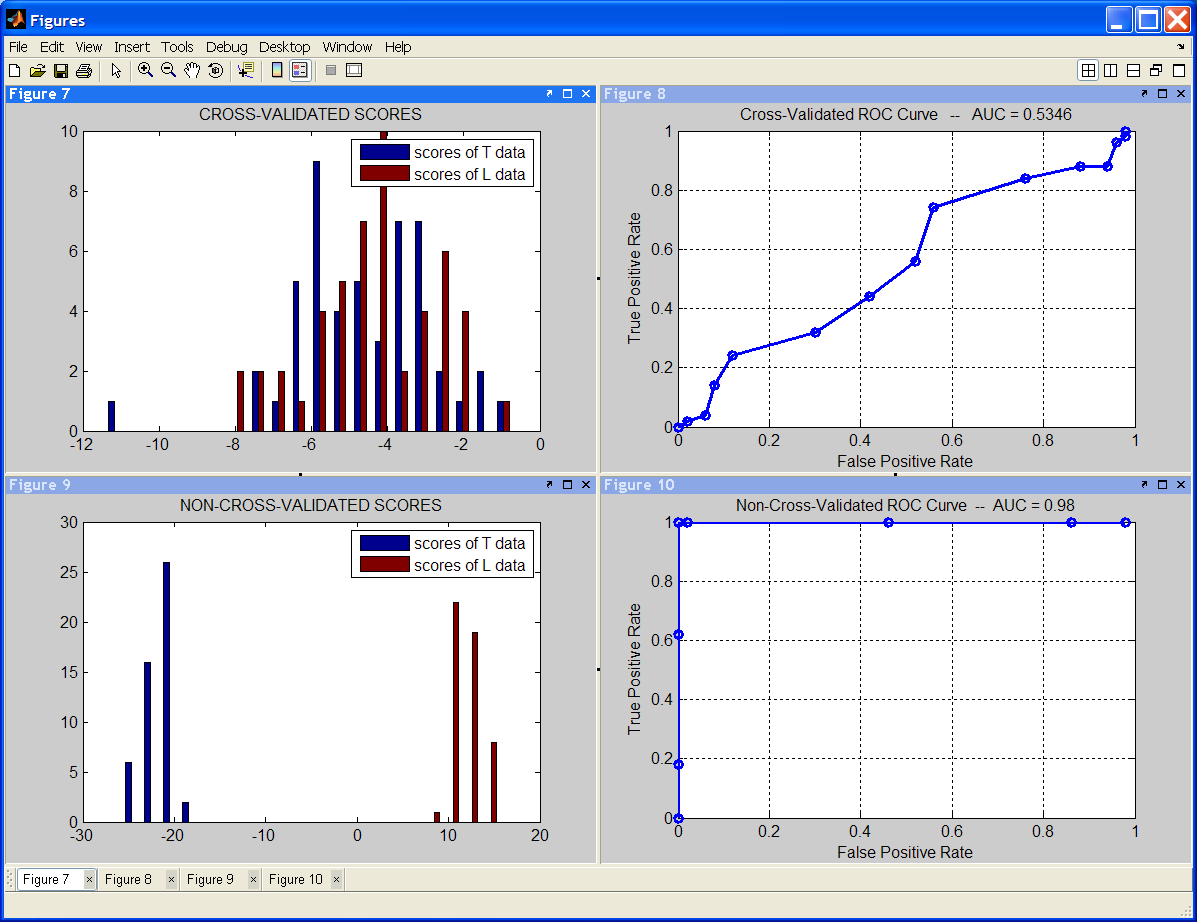
**Figure 2 Figure 3**

We will use two models: Model 1, and Model 2.

The scoring function (in both Model 1 and Model 2) does the following: for any given data image to classify, it takes the sum, over all voxels, of the data voxel intensity multiplied by the difference between the T Model and the L Model in that voxel.

Viewing each of the three images, namely the data to classify, the T Model and the L Model, as a vector of length 10,000, this is equivalent to taking the difference between the dot products of the data vector by the Model-T vector and the Model-L vector. Thus, if the data vector looks more like the L-model than the T-model, its score will be positive.

In Model 1, the estimated pattern is simply the sum over all training trials in each of the two conditions. No smoothing is used. The behavior of Model 1 can be summarized by saying that the model performs extremely well when tested on the same trials that were used to construct the model (the training trials), but performs extremely poorly when tested on new trials (the test trials). This is an unmistakable sign that Model 1 suffers from severe over-fitting. The reason is that no smoothing is used in Model 1.



**Figure 4. Cross-validation and ROC curves**

We see that the distributions of *cross-validated* scores in the two conditions are largely overlapping. The ROC curve ("Receiver Operating Curve," showing the rate of True Positives as function of the rate of False Positives for various values of the discrimination threshold D) constructed from these scores is close to the diagonal, and the AUC (Area Under the Curve, which should be close to 1 for a good model) is close to .5 (indicating that the model performs essentially at chance level). We conclude that there is no evidence, from the results obtained when using Model 1, that the data contains a T/L-related signal.

The distributions of the *non-cross-validated* scores in the two conditions are completely disjoint, and the ROC curve accordingly shows perfect performance, with an AUC of almost 1. This however is no evidence of the presence of a T/L-related signal, since these scores are computed exclusively from trials that were part of the training set, and this is "cheating."

In fact the comparison of the cross-validated results to the non-cross-validated ones shows beyond any doubt that Model 1 suffers from severe *over-fit*. In other words, the differences between the T-Model and the L-model are idiosyncrasies of the training trials, which don’t show up at all in the test trials.

We now turn to Model 2.

Model 2 follows a more elaborate procedure than Model 1: it smoothes the summed data by a bell-shaped filter. (Technically, it convolves the 2-dimensional summed-data image by a Gaussian-shaped “kernel,” of width ker\_width.) It then defines the model, i.e., the estimated activation pattern in each condition, as equal to the top 95th percentile of the smoothed image.

This operation of smoothing followed by thresholding has two effects:

1. It incorporates into the model, say the T model, voxels that don’t show up very strongly in the T-sum data, but are deemed to rightfully belong to the T pattern. The reason these voxels are deemed to be part of the T pattern in spite of not being strongly active in the T-sum is that, first, they are close to voxels that did light up strongly in the T-sum, and, second, **we know that, in principle, brain activation is a smooth function of space**. It is only because of the high level of noise inherent in the recording technique that these voxels did not light up strongly in the T-sum.
2. Similarly, voxels that are outside of the bright regions, but happened to light up strongly in the T-sum because of the noise, are excluded. This results in further noise reduction, hence, hopefully, in further increasing the signal/noise ratio.

When we set ker\_width to 2, we observe that the smoothed training-data sums (Figure 5 below) are still quite noisy, but are much smoother than the original training-data sums (Figures 2 and 3 above).

Model 2 (shown in Figure 6 below) show the result of thresholding the smoothed training-data sums at an intensity value corresponding to the 95th percentile. Thus, only 5% of the voxels are lit up. Model 2, unlike the smoothed or unsmoothed training-data-sums, is a binary-valued image (one for each of the two conditions).

From the cross-validated ROC curve, we see that if we set the discrimination threshold at a value such that the False Positive Rate is 0.5, we get a True Positive Rate of about 0.74. Equivalently, we have a True Negative Rate of 0.5 and a False Negative rate of 0.26.

With 50 trials in each class, this gives a total of 50×(0.5 + 0.74) = 62 for the number of correctly classified trials.

Under the Null Hypothesis that there is no T/L-related signal at all in the data, the classification algorithm is completely random. Therefore, the number of correctly classified trials ***X*** obeys a Binomial distribution with parameters 100 and .5. This distribution has a mean of 50 and a variance of 100×.5×.5, hence a standard deviation of ** = 5.

The observed value of our statistic ***X*** is more than two standard deviations away from the expected value under the null. Hence, based on Model 2 with ker\_width equal to 2, we can reject the Null Hypothesis that there is no T/L-related signal at all in the fMRI data.

The table below summarizes the results for increasing values of the kernel half-width, denoted *HW*. The Null Hypothesis is studied using a one-sided test, since we expect a value of the statistic ***X*** larger than the expected value ** under the Null. We report the ***Z***-statistic (also called *z*-score), namely:

***Z*** = (***X*** − ** )/** = (***X*** −  )/5

Under the null, *Z* obeys a standard normal distribution, hence we get the *p*-value for *z* (one-sided test) from a standard-normal table or by typing 1 – normcdf(z) in Matlab.

We also report the **Area Under the Curve** (**AUC**) and the **True Positive Rate** (**TPR**) that we get if we set the discrimination threshold at a value such that the False Positive Rate is 0.5.

***HW* *AUC* *TPR*** ***X*** ***Z* Reject H0 *p*-value**

0.1 .497 .53 51 0.283 No NA

0.5 .49 .55 52 0.40 No NA

1 .496 .50 50 0 No NA

1.2 .585 .65 57 1.40 No NA

1.5 .608 .67 58 1.60 borderline 0.055

2 .69 .74 62 2.40 Yes 0.008

3 .76 .82 66 3.20 Yes 7×10-4

5 .82 .91 70 4.0 Yes 3×10-5

10 .77 .88 69 3.80 Yes 7×10-5

12 .61 .67 58 2.0 Yes 0.023

15 .53 .56 53 0.60 No NA

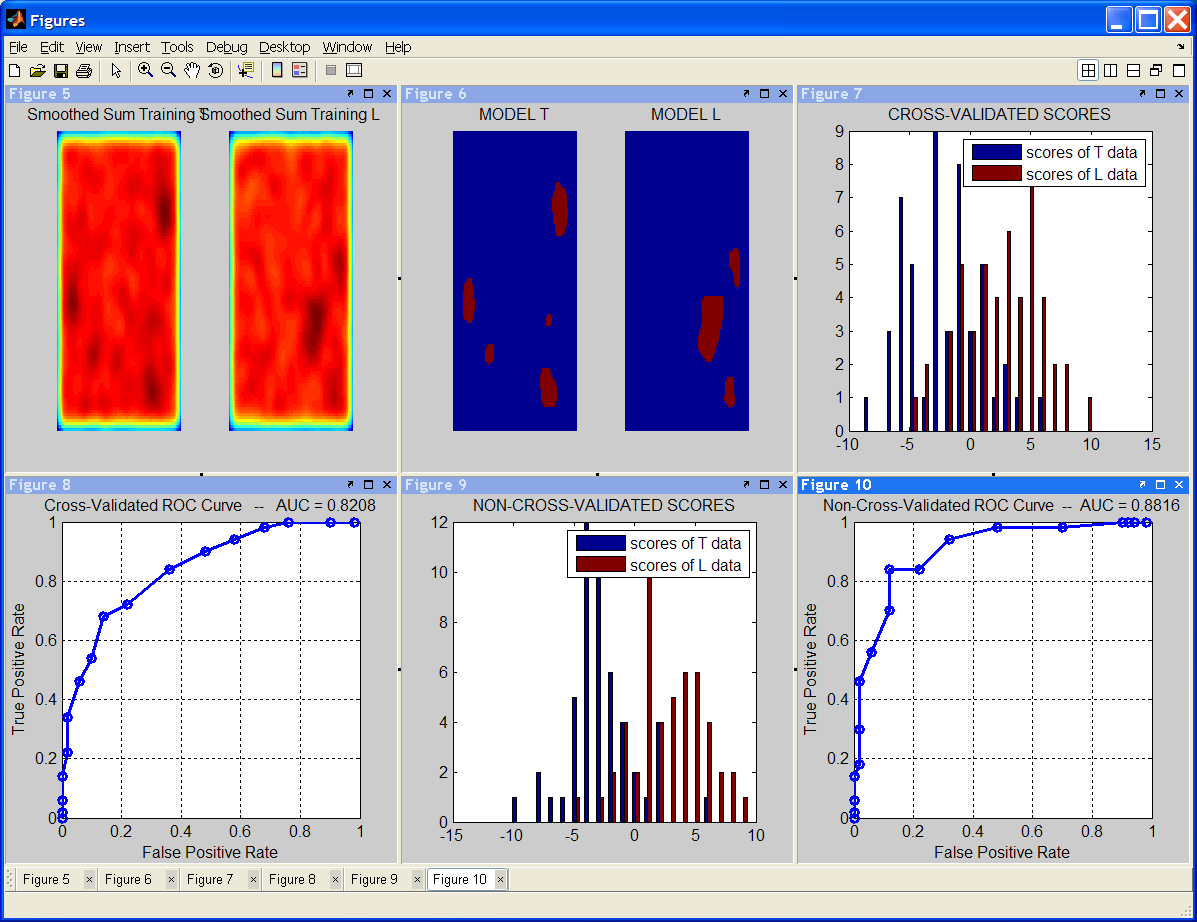
20 .44 .43 47 − 0.85 No NA

30 .58 .63 56 1.20 No NA

50 .494 .52 51 .283 No NA

In conclusion, the cross-validation procedure shows that the acceptable values for the smoothing parameter are roughly between 1.5 and 12, with an **optimal value of about 5**. This assessment is based on three criteria that stand in good mutual agreement: the AUC, the *p*-value when rejecting the null, and a qualitative inspection of the summed training data (Figure 5) and of the Models (Figure 6).

**Results for Kernel Half-Width (HW) = 5**



Visual inspection of the models shows the following behavior with increasing values of the smoothing parameter ker\_width (not shown here):

* a large number of “hot spots” of small size appear throughout the array;
* the largest hot spots gradually merge with each other while the smaller ones vanish;
* at the optimal value of ker\_width = 5 there are about 3 main hot spots in each model;
* the hot spots gradually move toward the center and merge into a single region, identical for the two models.

For very small values of ker\_width the model "over-fits" the data. For very large values it "under-fits" the data. There is a large range of acceptable intermediate values.