How many Principal Components?

L.K. Hansen, F.Å. Nielsen, P. Toft, C. Goutte — Department of Mathematical Modelling, Technical University of Denmark
S.C. Strother — PET Imaging Service, Minneapolis VA Medical Center, Minnesota, USA
N. Lange — Brain Imaging Center, McLean Hospital & Harvard University, Boston, USA
N. Mørch, C. Svarer, O.B. Paulson — Neurobiology Research Unit, Copenhagen University Hospital, Denmark
R. Savoy, B. Rosen — Massachusetts General Hospital, Boston, USA
E. Rostrup, P. Born — Danish Research Centre for Magnetic Resonance, Hvidovre Hospital, Denmark

474

474

Abstract

- Optimization of model complexity is the key to good generalization in exploratory statistical investigations in neuroimaging.
- We provide an operational tool for optimization of model complexity.
- The tool can be used to find the number of Principal Components or the number of clusters in Clustering.
- Illustrate the approach by Principal Component analysis of a fMRI experiment

Explorative statistics

Principal Component Analysis (PCA) and Clustering are two prominent examples of unsupervised learning schemes that are gaining interest in brain mapping.
Since learning proceeds from a finite sample of training data the estimated model parameters are stochastic variables depending on the particular (random) training set. Hence, there are questions of robustness and generalization: how robust are the estimated parameters to fluctuation and noise in the training set and how well will they perform on a new (test) datum?

Modeling and Generalization

It is a common observation that good generalization is obtained when the model capacity is carefully optimized:
If the model capacity is too limited the model will not be able capture the full complexity of the distributions, while a high-capacity model will support many different solutions to the learning problem and is likely to focus on non-generic details of the particular training set (overfitting).

Technicalities

A training set is a finite sample
$D = \{x_\alpha | \alpha = 1, ..., p\}$
of the stochastic measurement x.
Let $p(x)$ be the "true" distribution of x, while the "empirical distribution", associated with D, be $p_e(x) = 1/t \sum_{\alpha=1}^{p} \delta(x - x_\alpha)$.
For a specific model with associated error function $d(x|\theta, H)$ (e.g., the negative log-likelihood) and a specific set of parameters θ we define the training and generalization errors as follows,

$$E(\theta, H) = \int dx\, p_e(x) d(x|\theta, H) \quad (1)$$

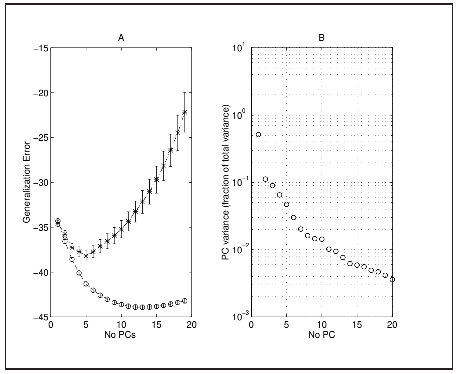
$$G(\theta, H) = \int dx\, p(x) d(x|\theta, H) \quad (2)$$

Note that the generalization error is non-observable, it has to be estimated either from a finite *test set* drawn from $p(x)$, or estimated from the training set using statistical arguments.

Quantification of generalization

The generalization error – involving the average on an infinite data set – can not be measured directly but must be *estimated*.
Here we investigate estimates based on

- A finite test set of data that are not used in the estimating model parameters ($*$).
- An analytical estimate based on asymptotic statistical sampling theory ($\circ$).



Typical generalization trade-off curve (left panel) used to optimize the number of Principal Components. For low-dimensional PC projections we underfit while for high-dimensional projections we overfit the retained PC's. In the right panel we show for comparison the eigenvalue spectrum of the correspinding PC's.

Sampling theory

In particular we assume that the measurement x (e.g., an fMRI slice with number of voxels L) can be modeled as a multivariate Gaussian variable s whose main variation is confined to a subspace of dimension K, this component being degraded by additive, independent isotropic noise n:

$$x = s + n \tag{3}$$

Hence, the "signal" $s \sim \mathcal{N}(x_0, \Sigma_s)$, while the "noise" is distributed $n \sim \mathcal{N}(0, \Sigma_n)$.
We assume (measurement model):

- Σ_s is singular, i.e., of rank $K < L$,
- $\Sigma_n = \sigma^2 I_L$, where I_L is an $L \times L$ unit matrix and σ^2 is a noise level.

We estimate the parameters $\theta \equiv (x_0, \Sigma_s, \Sigma_n)$ on the training set by maximum likelihood.
Furthermore, within maximum likelihood estimation we get a asymptotic estimate of the test error

$$\langle G(\widehat{\theta})\rangle_D \approx \langle E(\widehat{\theta})\rangle_D + \dim(\theta)/N, \tag{4}$$

where $\langle E(\widehat{\theta})\rangle_D$ is the average training error, and where the dimensionality of the parametrization (θ) grows quadratically with the number K of PCs retained in the PCA see e.g., Ref. (1).

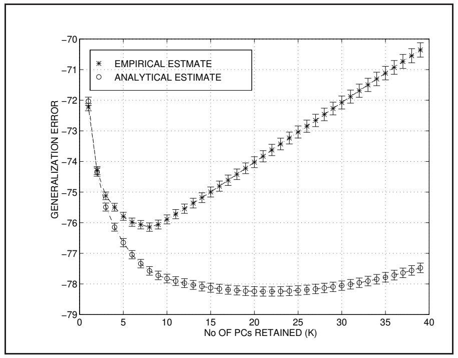
Experiments

The data has been collected with the following protocol:

subjects :	single subject.
task :	left-handed finger-to-thumb opposition.
modality :	whole-brain echo planar fMRI images.
paradigm :	multiple runs of 72 images. 24 baseline, 24 activation, 24 baseline
time step :	2.5 second.
location :	axial slice through primary motor cortex and SMA.
slice :	41x44 voxels.
voxel size :	3.1x3.1x8mm.

Complexity optimization

A training set of size (N =) 300 was drawn at random from the pool of scans, and another *disjoint* independent set of 324 scans were drawn at random to form a test set. PCA analyses were carried out on the training set. The generalization error was computed on the test set using the covariance structure estimated on the training sets and plotted versus size of the PCA subspace (K).



In the figure we show the progression of the generalization error as we increase the model dimensionality (K), as estimated on the test set $(*)$ and using the analytical asymptotic estimate (o). The error bars indicate the standard deviation observed in 10 replications of the experiment. Note that the minimal generalization error is obtained when projecting on $K = 8$ principal directions.

Discussion

It is apparent, from this example and indeed from many other similar examples we have investigated, that the asymptotic theory overestimates the optimal value of K and underestimates the overall the generalization error. Hence, we recommend the empirical test set-based scheme for estimating the proper dimension of the PCA subspace. It should be noted that principal component analysis is *unsupervised*, hence, does not use any information about the external stimulus (activation reference function). It can not be ruled out that interesting and generalizable structures are hidden in the rejected subspace, but such information has to be extracted using additional knowledge, e.g., about the stimulus.

Acknowledgement

This project has been funded by The Danish Research Councils and the Human Brain Project P20 MH57180 "Spatial and Temporal Patterns in Functional Neuroimaging".

References

1. L.K. Hansen, J. Larsen: Proc. Int. Conf. Neural Nets 1996