The Stability Principle for Information Extraction from Data

Bin Yu
Statistics and EECS, UC Berkeley

MMDS, Berkeley, June 22, 2016
Prevalence of False Scientific Discoveries

“A recent study* analyzed the cause of retraction for 788 retracted papers and found that error and fraud were responsible for 545 (69%) and 197 (25%) cases, respectively, while the cause was unknown in 46 (5.8%) cases.

— Casadevall and Fang (2012)

* R Grant Steen (2012), J. Med. Ethics

Casadevall and Fang called for

“Enhanced training in probability and statistics”
Scientific reproducibility

Statistical stability is a minimum requirement for scientific reproducibility

**Stability principle**: statistical conclusions should be stable to appropriate *perturbations* to data and/or models, and ad hoc decisions made by humans. Stability is defined relative to *stability measure(s)* corresponding to estimation, prediction or model selection, for example.
Case 1: Movie reconstruction using functional MRI signals

- **Non invasive and indirect recording technique**
- Low temporal resolution, a few seconds
- High spatial resolution
  - (voxel = 1x1x1 mm cube)
    - 10,000 voxels in early visual areas
    - each voxel covers > 100,000 neurons

- Can watch videos inside the machine
Movie reconstruction

Nishimoto, Vu, Naselaris, Benjamini, Yu and Gallant (2011)
Interpretable models through stability

Key step in movie reconstruction:

forward predictive model via Lasso with features indexed by locations on input image/frame

The science lies in finding the features that might drive a voxel. That is, we want interpretable models and stability is a minimum requirement for interpretability.
Interpreting encoding models: spatial locations of selected features

Prediction scores on Voxels A-C are 0.72 (CV)

Application of Stability Principle: Which perturbation scheme? Which stability metric?
Data perturbation has a long history

“he (Tukey) preferred to rely on the actual batch of data at hand rather than on a hypothetical underlying population of which it might be a sample.”

- p. 1643, Huber (2002)
  “JOHN W. TUKEY’S CONTRIBUTIONS TO ROBUST STATISTICS”
Data perturbation has a long history (continued)

Jacknife

Sub-sampling:
... Mahalanobis (1949), Hartigan (1969, 1970), Politis and Romano (1992, 94), ...

Cross-validation:

Bootstrap:
... Efron (1979), Bickel and Freedman (1981), Beran (1984), Künsch (1989), ...
Perturbation argument is central to limiting law results

One proof of CLT (bedrock of classical stat theory):

1. Universality of a limiting law for a normalized sum of iid through a perturbation argument or Linderberg’s swapping trick,

   i.e., a perturbation in a (normalized) sum by a random variable with matching first and second moments does not change the (normalized) sum distribution.

2. Finding the limit law via ODE.


Limiting laws are stability results.
Perturbation argument is central (continued)

1. Recent generalizations to obtain other universal limiting distributions e.g. Wigner law under non-Gaussian assumptions, last passage percolation, and high-dim M-estimation (Chatterjee, 2006, Suidan, 2006, El Karoui et al, 2014…)

2. Concentration results also assume stability-type conditions

Previous applications of “Stability Principle” to model selection

Model section in low-dim regression:

Lasso or sparse modeling:
   Bach (2008),
   Meinshausen and Bühlmann (2010),
   Liu, Roeder and Wasserman (2010),
ES: Estimation Stability (Lim & Yu, 2016)

Given a Lasso smoothing parameter $\lambda$, data perturbation scheme as in CV (or dividing the data units into $M$ blocks)

Get Lasso estimate $\hat{\beta}_m(\lambda)$ for each block $m = 1, \ldots, M$, and form an estimate $X\hat{\beta}_m(\lambda)$ for the mean regression function.

$$\bar{\hat{\beta}}(\lambda) = \frac{1}{M} \sum_m \hat{\beta}_m(\lambda)$$

**Stability measure (new):** estimation stability (ES) of mean regression function

$$ES(\lambda) = \frac{1}{M} \sum_m \frac{||X\hat{\beta}_m(\lambda) - X\bar{\hat{\beta}}(\lambda)||^2}{||X\bar{\hat{\beta}}(\lambda)||^2}$$
ESCV: Estimation Stability + CV

**ES** aims at **estimation stability of reg. function**, while CV aims at prediction stability.

ES is the reciprocal of a test statistic for testing

\[ H_0 : X\beta = 0 \]

with the variance of the mean function estimator estimated by delete-d Jacknife.
ESCV (continued)

ESCV selection criterion for smoothing parameter $\lambda$:

Choose the $\lambda$ that is the first local minimum of $ES(\lambda)$ that is larger than the CV choice.

Applicable to smoothing parameter selection in Lasso and other non-parametric regression regularization parameter selections.
Back to fMRI problem: spatial locations of selected features

Prediction on Voxels A-C: CV 0.72, ESCV 0.7

Sparsity gain with minimal prediction loss

Prediction (correlation)

Non-zero predictors

Based on validation data for 2088 voxels
Case 2:

Local gene network construction using spatial gene expression data

**Model perturbation**: relative to different modes in a non-convex optimization

**Stability measure**: Amari-index of matching columns of two matrices.

Drosophila shares 75% genes with human and is a model organism to study diabetes for example.
Data collection procedure (1/2)

* Using dye chemistry to visualize spatial gene patterns

* Imaging under a microscope
Data preprocessing (2/2): much work

- Use Erwin's segmentation algorithm to cut out the embryos
- Transform the embryo into an ellipse using SPEX2 (Puniyani and Xing 2010)

- Stain extraction

A sample of resulting images

- Segmentation stripes
- Brain
- Ventral neurogenic region
- Mesoderm
- Ventral epidermis
- Hindgut
- Posterior midgut
- Pole cells
- Anterior midgut
- Foregut

Dictionary

Coef. Matrix

Image 1

coef. Image 2

coef. Image N

405 pixels

K entries
The gap gene network: genes interact locally in space

Segmentation (vertical “coordinate”): work of the gap gene network

Fruitfly embryo: segmentation

Human embryo: segmentation

Nobel Prize in 1995 in Physiology or Medicine for work on gap gene

Christiane Nüsslein-Volhard
Reconstructing local transcriptional networks using spatial Drosophila embryonic expression images

February 21, 2015

1 Learning principal patterns

It is observed by biologists that a complicated expression pattern can be decomposed into a set of simpler localized principal patterns (Yakoby et.al [6]). Serving as building blocks for other patterns, each principal pattern usually corresponds to a specific functionality. Thus the study of gene expression can be simplified to first studying those building block patterns, and then mapping functionality of the constituting building blocks back to the original genes. In this project, we discover that Nonnegative Matrix Factorization (NMF, Lee and Seung [3]) can be used to obtain a set of data-driven building block patterns.

1.1 Nonnegative Matrix Factorization: the formulation

Suppose we observe \( n \) signals with nonnegative values each of length \( p \):

\[
\mathbf{x}_i \in \mathbb{R}^{p+} \quad \text{for} \quad i = 1, \ldots, n.
\]

NMF finds a dictionary \( \mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_K] \in \mathbb{R}^{p \times K} \) under which the signals have nonnegative representations: i.e.

\[
\mathbf{x}_i \approx \mathbf{D} \mathbf{a}_i \quad \text{for a nonnegative vector} \quad \mathbf{a}_i \in \mathbb{R}_K^+.
\]

Mathematically, NMF solves the following optimization problem:

\[
\min_{\mathbf{D} \geq 0, \mathbf{A} \geq 0} \| \mathbf{X} - \mathbf{D} \mathbf{A} \|_F^2,
\]

where the weight for the \( i \)-th image \( w_i \) is defined as one over the number of replicates of the gene that corresponds to the \( i \)-th image. Notice that the above weighted problem can be rewritten as

\[
\min_{\mathbf{D} \geq 0, \mathbf{A} \geq 0} \sum_{i=1}^n w_i \min_{\mathbf{a}_i} \| \mathbf{x}_i - \mathbf{D} \mathbf{a}_i \|_2^2,
\]

Interpretable data representation via Nonnegative Matrix Factorization (NMF)

References: ... Lawton and Sylvestre (1971)... Lee and Seong (1999), ... (SPAMS) by Marial et.al (2010).

References: ... Lawton and Sylvestre (1971)... Lee and Seong (1999), ... (SPAMS) by Marial et.al (2010).
staNMF uses a new stability criterion for selecting the number of principal patterns

Let $C \in \mathbb{R}^{K \times K}$ be the cross correlation matrix for two dictionaries $D_1$ and $D_2$ in $\mathbb{R}^{405 \times K}$. Define a dissimilarity measure between $D_1$ and $D_2$ as

$$diss(D_1, D_2) = \frac{1}{2} \left( \frac{1}{K} \sum_j (1 - \max_i C_{ij}) + \frac{1}{K} \sum_i (1 - \max_j C_{ij}) \right).$$

Brunet et al, 2004 used stability to perform model selection of NMF for clustering purpose, instead of dictionary stability.

$K = 21$ is the number of patterns selected by our stability criterion
Learned 21 Principal Patterns via staNMF

Anterior:

Vertical:

Horizontal:

Posterior:

The four types of patterns work together to define a rough “coordinate” system in early Drosophila embryonic development.
Principal patterns (PP) are biologically meaningful.
Distributions of local correlations for two networks

The distribution is bimodal: the negative bump corresponds to mutual repression and positive bump corresponds to mutual activation. Also, note the asymmetry of the distribution – tends to skew towards the positive direction. The mean is greater than zero.
Well-known Gap gene network: 10 out of 11 links reproduced

- Interactions correctly predicted
- Interactions that we missed

Anterior

Posterior

Repression
CRISPR/cas9 to knock-out CG-x

CRISPR/cas9
-- borrowed from bacteria immune system
(clustered regularly interspaced short palindromic repeats)

The CRISPR/Cas9 system “acts as a kind of molecular scissors, cutting and replacing DNA letters in an organism’s genome with exquisite precision and ease.”


J. Doudna  E. Charpentier  F. Zhang

https://www.brandeis.edu/now/2014/october/gabbay-award-2104.html

--http://en.wikipedia.org/wiki/CRISPR
Preliminary results on ftz after knocking out CG-x

fushi tarazu (ftz)
wild-type (normal)

CG-x CRISPR mutants
Preliminary results (continued)
Summary

• Stability principle is a minimum requirement for scientific reproducibility. It is intuitive and easy to use.

• We have shown two successful cases. Other applications also have good results.

• On-going and open problems:
  - theories for ESCV and staNMF
  - data perturbations and stability metrics in dependent case
Interdisciplinary Team on Movie Reconstruction Project

S. Nishimoto  J. Gallant
A. Vu  T. Naselaris

Y. Benjamini

ESCV Project

Funding: NSF, Guggenheim, NIH
Data Science Team on Drosophila Project

Siqi Wu  A. Joseph  K. Kumbier  S. Balakrishnan

Erwin Frise  A. Hammonds  S. Celniker

Wei Xu  Y. Zhang  F. Xu  A. Yao

Statistics, UCB

Tsinghua Univ,  IIIS

Funding: NSF, NIH
References

Bernoulli (2013)

*Bernoulli* 19(4), 2013, 1484–1500
DOI: 10.3150/13-BEJSP14

Stability

BIN YU

JCGS (2016)

Estimation Stability With Cross-Validation (ESCV)

Chinghway Lim and Bin Yu

Current Biology (2011)

Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies

Shinji Nishimoto,1 An T. Vu,2 Thomas Naselaris,1 Yuval Benjamin,2 Bin Yu,1 and Jack L. Gallant1,2,A
1Helen Wills Neuroscience Institute
2Joint Graduate Group in Bioengineering
3Department of Statistics
4Department of Psychology
University of California, Berkeley, Berkeley, CA 94720, USA

mental processes. It has therefore been assumed that fMRI data would not be useful for modeling brain activity evoked during natural vision or by other dynamic mental processes. Here we present a new motion-energy [10, 11] encoding model that largely overcomes this limitation. The model separately describes the neural mechanisms mediating visual motion information and their coupling to much slower hemodynamic mechanisms. In this report we first validate this an...

PNAS (2016)

Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks

Siqi Wu1,2, Antony Joseph3,4,5, Ann S. Hammonds2, Susan E. Celniker1, Bin Yu6,7, and Erwin Frise1,8
1Department of Statistics, University of California, Berkeley, CA 94720; 2Division of Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA 94720; 3Walmart Labs, San Bruno, CA 94066; and 4Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720

6/21/16