Dimension Reduction

David Hofmeyr

Dept. Statistics and Actuarial Science,
Stellenbosch University, South Africa

6-8 December 2021

UNIVERSITEIT
iYUNIVESITHI
STELLENBOSCH
UNIVERSITY
100
1918·2018

forward together · saam vorentoe · masiye phambili

Outline

Overlearning

Model Selection and Validation

Random Projections

Overlearning

- Essentially the same as overfitting, but for unsupervised context
- Allowing small (natural) sample variations to accumulate and dominate the "learning"
- With enough degrees of freedom we can make data look however we want
  - If $p >= n$ then for ANY $z \in \mathbb{R}^n$ there is a $\mathbf{v} \in \mathbb{R}^p$ s.t. $Xv = z$ (assuming full rank X)
- Even in a non underdetermined system we can still do a lot of harm...

Overlearning

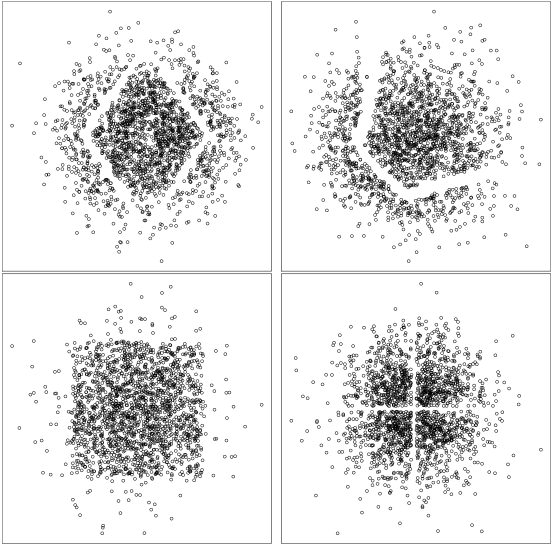

Figure 1: Projections of the SAME DATA simulated from a GAUSSIAN distribution

Model Selection and Validation

- So how can we trust what we “see” in a projection/visualisation of data?
- How can we measure degrees of freedom, or some other form of “model” complexity?
- It is very telling of the difficulty that the earliest record of some concept of degrees of freedom for even standard PCA (i.e., the vanilla homoschedastic Gaussian residual model) is only about 10 years old.
- This is my favourite topic these days, but is extremely challenging

Model Selection and Validation

- One approach: Model “sensitivity” through leverage
  - Model complexity measured as

$$\sum_{ij} \frac{\partial}{\partial \mathbf{X}_{ij}} \left(\mathbf{X} \mathbf{V}^{*} \mathbf{V}^{*'} \right)_{ij}$$

  - Equivalent to effective degrees of freedom in homoschedastic Gaussian error model
- This is hard:
  - Differentiating optima is hard enough
  - differentiating *constrained* optima is harder

Model Selection and Validation

- For the first column of V^*, using the PP formulation, we get

  $$n - (p-1) - 2\nabla_{\mathbf{p}}\phi(\mathbf{p}^*)'\mathbf{p}^* tr(H_\Phi(\mathbf{V}^*_{:1})^+),$$

  where $\mathbf{p}^* = \mathbf{X}\mathbf{V}^*_{:1}$, H_Φ is the Hessian operator of the objective and $^+$indicates the Moore-Penrose Pseudo-inverse

  - This one is easier because it doesn't need orthogonality constraints
  - For constant linear orthogonality constraint: $\mathbf{V}^{*'}_{:k+1}\mathbf{A} = \mathbf{0}$ we get

    $$n - (p-k-1) - 2\nabla_{\mathbf{p}}\phi(\mathbf{p}^*)'\mathbf{p}^* tr((\mathbf{F}' H_\Phi(\mathbf{V}^*_{:1})\mathbf{F})^+),$$

    where k is the rank of $\mathbf{A}$ and $\mathbf{F}$ has as columns an orthonormal basis for the null space of the columns of $\mathbf{A}$.
  - BUT our constraints are NOT constant, they're given by $\mathbf{A} = \mathbf{V}^*_{:[k]}$

Model Selection and Validation

- The formulation where the fact that **A** changes with the entries in **X** involves a hideous expansion which I haven't been able to simplify
- BUT, in the standard PCA formulation some nice cancellation occurs, and the result agrees with prevailing theory, which is very encouraging
- The only other ideas I have had involve PAC bounds on the empirical distribution of **XV** from VC theory and Rademacher/Gaussian Complexity, but these are very loose bounds and the theory I find very tough

Random Projections (RPs)

- Yup, that's pretty much it!
- Under very mild assumptions, if the entries in $\mathbf{V}$ are random, you can maintain structure from $\mathbf{X}$ very well
- Some properties of RPs
  - (scaled) RPs $\approx$ orthonormal
  - Pairwise distances $\approx$ preserved under RP!
  - "Typical" RPs of data are close to Gaussian (marginally)
  - Some ML/statistical methods have rigorous performance guarantees under RP: Random $\mathbf{V}$ gives $\approx$ sufficient dimension reduction

Random Projections

- example: $vec(\mathbf{V}) \sim N(\mathbf{0}, p^{-1/2}\mathbf{I})$
  - $\mathbf{V}$ is (approximately) distributed uniformly on the Stiefel Manifold
  - $\approx$ Orthogonality follows from $\mathbf{V}'_{:i}\mathbf{V}_{:j} = \frac{1}{p}\sum_{k=1}^{p} Z_{ki}Z_{kj} \xrightarrow{a.s} \mathbf{0}$, since $E[Z_1 Z_2] = 0$, $Var(Z_1 Z_2) = 1$ for independent standard normal Z_1, Z_1
  - $\approx$ Unit norm follows from $||\mathbf{V}_{:i}||^2 = \frac{1}{p}\sum_{k=1}^{p} Z_{ki}^2$ with $\sum_{k=1}^{p} Z_{ki}^2 \sim \chi_p^2 \Rightarrow E[\sum_{k=1}^{p} Z_{ki}^2] = p$, $Var(\sum_{k=1}^{p} Z_{ki}^2) = 2p$, and so

$$\frac{\sum_{k=1}^{p} Z_{ki}^2 - p}{\sqrt{2p}} \approx_P 1 \Rightarrow \left|\frac{1}{p}\sum_{k=1}^{p} Z_{ki}^2 - 1\right| \approx_P \sqrt{2/p}$$

Random Projections

- **Theorem:** (Johnson and Lindenstrauss Lemma, "paraphrased") For $0 < \epsilon < 1, n \in \mathbb{N}, p' \geq 4\log(n)(\epsilon^2/2 + \epsilon^3/3)^{-1}$, for any set of n points in $\mathbb{R}^p$ and random projection $\mathbf{V} \in \mathbb{R}^{p \times p'}$ and with probability at least $1 - 1/n$

  $$(1-\epsilon)\sqrt{p'/p}d(\mathbf{x}_i, \mathbf{x}_j) \leq d(\mathbf{V}'\mathbf{x}_i, \mathbf{V}'\mathbf{x}_j) \leq (1+\epsilon)\sqrt{p'/p}d(\mathbf{x}_i, \mathbf{x}_j),$$

  simultaneously for all i, j, where d is the Euclidean metric.

Random Projections

- **Theorem:** (S. Dasgupta)
  X a random variable on $\mathbb{R}^p$ with zero mean and finite second moment. Let F_R be the distribution function of $\mathbf{V}'X$ with $\mathbf{V}$ as we had before. Then for any measureable B define

$$F(B) = \int_{\mathbb{R}} \int_B \frac{1}{\sqrt{2\pi\sigma}} \exp(-||\mathbf{y}||^2/2\sigma) d\mathbf{y} g(\sigma) d\sigma,$$

where g is the pdf of $||X||/\sqrt{p}$. Then

$$P\left(\sup_{\mathbf{x}\in\mathbb{R}^{p'},\epsilon>0} |F_R(B_\epsilon(\mathbf{x})) - F(B_\epsilon(\mathbf{x}))|\right) \le K,$$

where K is dominated by $(\frac{{p'}^2}{p} \frac{\lambda_{max}(\Sigma_X)}{G^{-1}(\epsilon)^2})$

Random Projections

- Does this result ruin everything?
- If $\frac{\lambda_{max}(\Sigma_X)}{G^{-1}(\epsilon)^2}$ is small then "sort of"
- Example: $X \sim 0.5N(\mu, \mathbf{I}) + 0.5N(-\mu, \mathbf{I})$ with $\mu_i = \theta$ in p_1 locations and zero otherwise.
  - $\lambda_{\text{max}}/G^{-1}(\epsilon)^2 \frac{p_1^2}{p} \geq\approx \frac{p_1^2+\theta^2 p_1^3}{p+p_1\theta^2}$

References

High Dimensional “issues”

- Beyer et al., When Is “Nearest Neighbor” Meaningful?, *Proc. Database Theory — ICDT’99*, 1999, Springer
- Some nice content in Wegner, S-A. *Lecture Notes on High-Dimensional Data*, `https://arxiv.org/pdf/2101.05841.pdf`

Projection Pursuit

- Theory/Overview: Huber, P. Projection Pursuit, *Annals of Statistics*, 1985
- Practical/Implementation: Hofmeyr, D. and Pavlidis, N. PPCI: an R Package for Cluster Identification using Projection Pursuit, *R Journal*, 2019
  Hofmeyr, D. Fast Kernel Smoothing in R with Applications in Projection Pursuit, *JSS* to appear, also `https://arxiv.org/abs/2001.02225`

References

Independent Components Analysis

- Hyvärinen, A., Oja, E. Independent component analysis: algorithms and applications. *Neural networks*, 2000

Random Projections

- Johnson, W., Lindenstrauss, J. Extensions of Lipschitz maps into a Hilbert space, *Contemp. Math.*, 1984
- Dasgupta, S., Gupta, A. An elementary proof of the Johnson-Lindenstrauss lemma, *International Computer Science Institute, Technical Report*, 1999
- Dasgupta, S. A concentration theorem for projections, *Proc. UAI*, 2006
  - Aside: If Sanjoy Dasgupta has done work in your area, he's probably done some of the best of it. There are tons of people who've done much more excellent work here, but he is a personal hero of mine