3 books with Tibshirani (bible, R, and sparsity)

Lasso

Murphy – probabilistic presentation

Bishop

Dubois – pattern recognition

Benjo – deep learning

Factor models – revise

Project

A small paper (approx 8 pages)

Ex: some existing estimator is not about heterogeneity, but about non-linearity in the main regressor

Need to be sure that the method works/helps – clear thinking and discussion – simulations (data driven; not too artificial) – tuning (be aware of necessary choices; cross-validation is one way; choice of tuning can affect the result significantly ?? what's the point of the choice then)

Theory – explain how your method works

probabilistic vs statistic ???

2 cultures:

Model vs black-box

if black-box – focus on the quality of prediction – out-of-sample – training sample (estimation sample) vs testing sample – an important shift in the measure of success (not theoretical properties but real-world performance)

Daniella Witten – a new paper about redux – point out lack of transparency in black-box approach, need to interpret what's happening inside ML estimation – recognise the 'assumption that we need' ???

Big data – principally different structure of data (text, networks, etc. – time series??) – big data is a collection of small datas – but need to recognise the hierarchical structure – within vs between analysis – ex: when learning about a particular firm can use info about other firms (which is helpful) – naturally, big models are necessary which means a lot of parameters (can we treat 'complex' data with small number of obs as big data?) –

natural concern is overfit as a result

Overfit: fitting noise rather than signal – ex: too many instruments makes 2sls useless (bc predicts the error rather than x) – sol: regularisation (accepting bias in the estimate to avoid fitting the noise); ex: lasso

Note about training/test samples: existing testing samples are too 'close' to the training, while economics is interested in very 'distant' samples – that's why we need models

OH qs

- S mentioned the analysis of consistency of these methods – but are identification issues any different in big-dimension contexts?
- is black-box similar to just being fully non-parametric? or is it also about assuming away treatment assignment issues?
- big data – is the number of obs a necessary condition – are dbns a complex enough object to be classified as big data? – want to clarify the ex with firms
- probabilistic vs statistical – what's the meaning?

lecture 2

$$T = \frac{NT}{N} = \frac{obs}{units} \tag{1}$$

Example where the variance doesn't go to 0 in asymptotics

$$y_{it} = \rho y_{it=1} + \alpha_i + u_{it} \tag{2}$$

Class 4: Nonparametric regression

$$y = f(x) + u\mathbb{E}[x'u] = 0 \tag{3}$$

Basis expansion:

$$\sum \beta_k \phi_k \tag{4}$$

keeping the size of the series fixed – in asymptotics just get the minimizer of the expected error – in other words, the bias is high due to approximation error

Can we confidently control where on the bias-variance tradeoff we are

High-dimensional regression

$$y_i = \sum_{i=1}^{p} \beta_i x_i + u_i \tag{5}$$

$$p = p(N) \tag{6}$$

Confidence intervals (for what ??) need to account for this weird asymptotics

what is consistency for basis expansion estimators? particular values of the target function?

Lecture 3

PCA – some of the beta are gonna be better estimated than others (is that what stephane said?)

New asymptotics – number of parameters grows with the sample siez

What is the second method ?

What is the meaning of MSE when you do prediction ?

Does lasso estimate directional derivatives well ?

Robust statistis

Doing inference for the prediction?

The idea behind the new asymptotics – the challenge of estimation is that the number of parameters is too large compare to the number of obsrvations

Would the hubert results hold if i take differences instead of a prediction

Hubert deals with the case when there's no exogeneity ??? if there is exogeneity – the basic results hold ???

Approximation with a series of functions – 2 things: the number of elements of a series and how well the target function can be approximated ?

Non parametric estimation – smoothness

Lecture 4

OLS – estimates the best approximation to what ? Now look at a more complex approximation Are there 'uniform efficiency' results?

Nuy (1997)

Euclidean norm (mean squared error) is not the right objective function in polynomial approximation (overfitting concerns)

Minimise the difference with the signal Does it change things for the linear case (compared to OLS)

The rate of convergence – N ?

Can we make inference on the complexity of the true function? Is there a possibility that we 'overdo' – take too many terms

Decomposing problems of approximation and of sampling.

Approximation error – are there different difficulties with multiple regressors – do people do approximation of derivatives

Does this analysis give out an optimal number of approximation terms to include?

Does it ensure that the optimal K is less or equal than the true number of terms.

loose upper bound if we know k_0 – the objective function is intended for approximation error being there all the time – this upper bound is a bad approximation to the mean squared error

cross-product between sampling error and approximation error

matrix $K \times K$ – $\phi'\phi$

spectre does not go to zero ??

oh with stéphane april 12

can lasso be interpreted as credibly estimating certain marginal effects

mixed integer programming – how can you get a certificate saying that the solution is global

rdd – running variable is also estimated – sample correction how to smooth things out in discrete outcomes settings – Manresa pouliot kaji

generalised indirect inference – how to deal with discrete problems

correlation neglect

salaix martin – 2 mln regressions cross-country national discord

panel data / group data – functions of parameters are still estimated okay

efron large scale inference

Lecture 6

Subset selection – showing that the probability choosing a wrong j is low

Union bound:

$$P\left[\cup_{k=1}^{K} A_k\right] \leq K sup_{k\in\{1,\dots,K\}} P\left[A_k\right] \tag{7}$$

Can we choose optimal s based on this criterion?

Finite sample ?? Is there a similar result for asymptotics for arbitrary dbns?

Show that the tail of a normal pdf shrinks exponentially

Can you use the the existing sample to infer the rate of decline of $\log p/N$

Markov inequality – can we at least expand the range of cases for which we can get bounds

Exponential bounds – a) dependence structure; b) tails

Asymptotics are hard because p is very large relative to N – Chernozhukov works on developing tools here

Lasso

What is a sparse DGP?

Idea:

$$min||Y - X'\beta|| \tag{8}$$

$$b \to ||b||_1$$
$$\mathbb{R}^p \to \mathbb{R}_+$$

Tibshirani paper

Lasso – both shrinkage and thresholding

Is this still backward induction type of approach ?

Does lasso work well only for sparse DGPs? Or sparsity is just imposed in the estimation process but we're still estimating an 'arbitrary' DGP?

Lecture 7

When choosing different λ in lasso, do we estimate different features of the joint dbn of Y and X ? or the same one, but we're configuring the best esimation strategy ? The choice of λ definitely affects the rate of convergence - but this is an estimation property, right?

OLS's rate of convergence with large p is not good – but don't we also estimate a different thing ??

non-multicollinearity assumption is hard in the large p setting – we could restrict it to the s actually 'non-zero' regressors – but we don't know this actual subset – so we come up with an alternative assumption that is convoluted and tailored to this setting – minimal eigenvalues

minimal eigenvalue condition

is the key (relaxed) condition essentially a restriction on the ratio of numbers of zero and non-zero elements in the parameter vector ??

What is the relationship between the optimal λ and the minimum λ that would yield all non-zero parameters ?

Lecture 9

Support recovery – whether we manage to obtain actually important coefs?

Sudden q: Why do we also need shrinkage beside thresholding? Is it just a side effect of better thresholding?

Why does OLS on pre-selected good regressors yields the consistent estimate ? How does shrinkage happen ?

So, shrinkage is not something that we actually need; i.e., it's not something that we care about in terms of the model – it's just an estimation feature ?

moment inequalities as an approach to do inference on model selection – relation to ddml where we care about only one parameter so don't need to do inference on model selection ??

Lab and Pischa (2005, ET)

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + u_i \tag{9}$$

Trade-off between noise coming from model selection and noise coming from correlation between x_1 and x_2 (in estimating β_1

Approach – introduce the concept of a threshold for β_2 to decide if it's important

recall heckman pset 1 ???

Compare to Hausman test (OLS vs IV); RE or FE; weak IV and F-stat; information criterion the idea of ddml is to avoid this issue basically

I thought the logic is: include control – estimate didn't change – why lose on inference then ?

Side note: RE vs FE – how does the interpretation of coefs change ???

Can i use bayesian – and offer two inferences on the same coef conditional on two different models being true

The idea/solution – the estimator is a mixture of two estimators

Is it gonna be structurally different if i assume the opposite – β_2 starts with 0 and converges to some non-zero value apparently converging to 0 is an important distinction – signal to noise does not improve with the sample size – alternative: blow up the variance of u with the sample size (rather than make β_2 go to 0)

solution: use the long regression – but that changes the interpretation of the parameter also !!!!

Approach 2: approximate the long regression frisch-lowell

$\hat{\beta}_2$ is a function of both the role of x_2 in the long regression and correlation between x_1 and x_2 – ddml is about separating these two ?

orthogonalisation

when ddml does badly, can it be so bad to be worse than the naive approach ??

how did we circumvent impossibility

duflo, chernozhukov 2018

Lecture 10

Belloni

Does lasso + basis expansion give a better estimate for cef derivatives ?

Newey - series expansion and average derivatives

you need some smoothness (of both approximand and approximator) for these approximations of derivatives to work well

Basic nonparametrics

What's the point of making the distinction about kernel estimator treating functions to be different at each x?

Nadaraya-Watson

Local linear regression

$$\hat{f}(x) = \frac{\sum_{i=1}^{N} \mathbf{1}\{||x_i - x|| < \varepsilon\} y_i}{\sum_{i=1}^{N} \mathbf{1}\{||x_i - x|| < \varepsilon\}} \tag{10}$$

K-nearest neighbours

$$\hat{f}(x) = \frac{\sum_{i=1}^{N} \mathbf{1}\{x_i \in A\} y_i}{\sum_{i=1}^{N} \mathbf{1}\{x_i \in A\}} \tag{11}$$

Should we treat the set of chosen neighbours as hyperparameters? That we can check using cross-validation ??

The trade off between using 'more data' (like using y besides x) and the complexity of the obtained estimand (in terms of interpretation) – a big topic in ML

Binning – partition $suppX$ (globally)

$$\hat{f}(x) = \tag{12}$$

K-means clustering

Maximise across-cluster variance and minimise within-cluster variance

Correspondence between PCA and K-means clustering – spectral clustering of networks (step 1 – do K-means to principal components)

Regressing what on what to reflect the K-means logic

Metrics that are immune to arbitrary rescaling of data ??

The objective function of the K-means:

Can we choose to optimise over 'planes' that dissect

Lecture 11

Honesty principle ??

Is this the right way to think: cluster X – within each cluster the dbn of Y is different

Can there be weird partitions with disconnected subsets ?

Does a tree do better than k-means with cross-validation (or some other tuning)?

Can we think of all this partitioning as finding conditional averages of the function? Does this interpretation lead to important differences for the method ?

When unrestricted partitions - we 'forget' about X – wow this seems to be another take at the 2 dimensions problem (in my world lol)

symmetric and cyclical functions ?? might still be useful to let disconnected be ??

Approach: target only simple functions defined on connected rectangles

rectangles are only vertical/horizontal lines ?? diagonals don't count? what about linear pivots ??

VC class !!! – partition the function space ???? according to complexity of the function ???

entropyyy !!!!!!!!!!!!!!!!

Myungkou's oh – may 5

What does lasso estimate if the model is not really sparse ? This is so relevant to my ongoing quest to try to understand the difference between making assumptions and just estimating some weird parameter.

Take some scalar rv y and an infinite sequence of some vars (just an infinite vector) $\{x_k\}_{k\in\mathbb{N}}$.

Need a continuum of variables, no??

How to shoot to infinity that is continuum ?

pca + lasso

TA session – may 9

inference for out-of-sample predictions

Lecture 12

imposing complexity constraints in the sake of reducing variance - - variance around what ?

CART is a 'stepwise' approx to the actual joint problem. Are there any bound guarantees on the error?

can you combine cart with kernelling? e.g., choose kernel types depending on the data ???

is the algo polynomial in the number of covariates?

the analogue of the coordinate-wise algo?

is this actually pure model selection? seems like choosing rectangles is the actual model selection to me ??

is there correspondence between cart and basic k-partitioning? i.e., are there simple cases when they give the same output? or correspondence with kc-means?

when choosing the next leaf – maybe take into account the variance of X too? to kinda anticipate the reduction in the MSE?

if ur interested in prediction for just a particular small subset of X, is there any sense in starting CART with the whole space ? can we modify the algo to improve in terms of variance ??

pruning - - this is just a particular way to choose subsamples ? what about bootstrap ?

big big big big issue of how to differentiate between variation across domain and across sample space – takes an interesting angle in this context

bayesian updating of the complexity of the model - - sounds cool

forests – want to reduce the variance

two ideas:

- bootstrap full sample with replacement (bagging) – but what about bootstrapping SUBsamples; shouldn't that move us closer to independence of draws ???
- randomly subset covariates on the second stage

when people say "work well in applications", do they mean that simulations of angrist type worked well??

Lecture 14

Honest inference

What about the F-test with point=wise inference ? Why is it worse than uniform bounds?

Necessary condition to test linearity (recall my idea some time ago)

Uniform bounds are kinda a continuous version of the F-test, apparently.

Idea: Thomas said that trees work badly for categorical vars; can the reason be that we still use the usual euclidean metric for partitioning?

Wager & Athey – paper about honest inference

Is it useful to think in mixture/hyperparameter way about the distribution of z_i? We first choose from what subpopulation we draw, then we make a draw (just a single one, basically)??

undersmoothing vs overfitting (focus on different aspects of the b/v trade-off ?)

ass: tree is deep enough so that bias is relatively less important than variance

sample splitting

my thinking: introduce a dbn over partitions – then we first pick a partition, then realisations

relation to ddml? simplify theory comment; ddml does not necessarily require sample splitting

fit the partition on one subsample, compute average on another subsample

how to do sample splitting for dependent data

trade-off between low bias obtained with finer partitioning and high variance of the mean of y in a leaf

cross-validation idea to find the optimal split size

neural net - - would it help to focus on the gradient ?? i.e., a differential form ?

Lecture 15

series approximation – works well only locally – why can't we use different points as approximation points ? is this kernelling ?

Neural nets

Approximation as the number of nodes increases

linear approximation – error in the limit – what about local approximation ?

asymptotics vs finite sample - - in kernel, bandwidth doesn't show up in the limit , is that the point ? - - so need to look at finite sample properties ?

for which dgps do different methods work better

NLLS

objective is non-convex

are the nice properties of newton-raphson dependent on convexity of the obj?

back-propagation aka "chain rule"

numerical approach is bad bc calculation of derivatives is too costly

stochastic gradient descent

choice of μ – is it only about speed, or for some μ there'll be no convergence in the limit deep net – more parameters make computation easier. does this have consequences for variance of the *estimator*

are local minima better than other non extremum points

multicollinearity

Lecture 16 – factor models

mapping economic models to factor analysis – see how dsge is mapped to vars

one of the ways to think: you want to use as much info as possible to infer some parameter – but it's not about making your sample larger, it's about using info from *other* sources.

is ridge in fact lasso over eigenvalues ? *Stéphane's answer:* maybe in Frobenius norm (?)

$$\begin{aligned} YV_1 &= [USV'] \, V_1 \\ &= US \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \\ &= s_1 U_1 \end{aligned} \tag{13}$$

DGP to approximate factor models (to introduce 'shocks' to the factor)

can we only use linear functions of factors (to define observables) ?

factor is just something that is supposed to predict what we want – so it's basically looking at what the important treatment is, no? – can i apply this logic to identifying important mechanisms (going back to problems that i've been having with my project)

Rotation problem – seems to be directly related to the issue of all causes vs potential outcomes !!!! – all causes leads you to this scaling issues – rotation thing is actually more general than than (but also includes scaling)

Choice of k determines how many first eigenvalues to take when doing PCA.

Asymptotic behaviour of eigenvalues and directions – is it conditional on the choice of k?

Asymptotics of PCA

Trade-off: learning more about λ vs f.

For forecasting: do time series analysis on factors – i.e., doing time series prediction on factors and then predicting the actual outcome of interest is preferable ? there must be a trade-off too bc of noise in estimating factors, right ?

Matrix penalisation

Least-squares estimator:

$$\min_{\lambda,f} \sum_{i,t} (y_{it} - \lambda_i' f_r)^2 \tag{14}$$

Equivalent to:

$$\begin{aligned} \min_{\lambda,f} \quad & \sum_{i,t} (y_{it} - m_{ir})^2 \\ \text{s.t.} \quad & M \text{ has rank} \leq k \end{aligned} \tag{15}$$

Problem: $rank(M)$ is not a smooth function of M

Put a constraint on the the sum of eigenvalues.

Nuclear norm of M:

$$||M||_* = \sum_{k=1}^{\min(n,T)} s_k \tag{16}$$

Because a norm is convex, we obtain a convex minimisation problem !!

Relation to lasso is interesting.

This does both shrinkage and thresholding to singular values.

Look at Wrainwright stuff.

Application 1

Interacted fixed effects

$$y_{it} = \beta' x_{it} + \lambda_i' f_t + \varepsilon_{it} \tag{17}$$

$$\min_{\beta,M} \sum_{i,t} \left(y_{it} - \beta' x_{it} - m_{it}\right)^2 + \lambda ||M||_* \tag{18}$$

Application 2

Matrix completion

$$\min_{M} \sum_{i,k} \left(y_{it} - m_{it}\right)^2 \tag{19}$$

Lecture 17 – latent vars

x_{it} is strictly exogenous

$$f(y_{i1}, \dots, y_{iT} | x_{i1}, \dots, x_{iT}) = \int f_y(y_{i1}, \dots, y_{iT} | a_i, x_{i1}, \dots, x_{iT}) \times f_a(a_i | x_{i1}, \dots, x_{iT}) \, da_i \tag{20}$$

Static model:

$$f_y(y_{i1}, \dots, y_{iT} | a_i, x_{i1}, \dots, x_{iT}) = \prod_{t=1}^{T} f_y(y_{it} | a_i, x_{it}) \tag{21}$$

Killing two birds: non-random treatment assignment & complex dependence structure. Is there a correspondence between these two goals?

world prices for a small country – example of exogenous x_{it} ?

kernel type of dependence? i.e., you're correlated with ‘close’ enough units

is it easy to estimate density of a_i if we have a balanced panel?

networks

is choosing a_i related to partitioning of nodes ?

can you impose high correlation for types conditional on being ‘close’ in the network? would that help? what i mean is that if someone draws a_i, that affects probabilities of drawing a_i for neighbours.

Pickel ?

Variational inference – approximation approach