What is the high dimension?

1. **Different tools and goals**
2. **Examples in Economics**
3. **Outline of the course and references**

Different tools and goals

Two cultures. Read Breiman (2001).

Ideas:

- Curse of dimensionality. Statistical and computational. Number of data points and number of parameters increase (and standard errors can still be large with big data).
- Regularization. Modify the estimator to make it well-behaved in some circumstances. Penalization ideas.
- Sparsity. Example of a linear regression:

  $$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + u_i,$$

  where only some of the β_j's are different from zero.

  We hope to exploit sparsity or structure. If nature is too complex there are sharp limits on what we can learn (Stone's optimal rates for nonparametric regression for example).
- Model selection.

  Select the set of relevant regressors in a sparse regression model.
- Model averaging.

  Average over possible solutions (e.g., obtained by selecting s out of the p regressors). How to weight the terms in the average?

- Adaptivity.

  $$y_i = \sum_{k=0}^{K} \beta_k \phi_k(x_i) + u_i,$$

  versus

  $$y_i = \sum_{k=0}^{K} \beta_k \phi_k(x_i; \theta_k) + u_i.$$

  In adaptive methods we choose the model in a data-driven way. A key feature of most ML methods (trees, neural networks, kmeans...) is their adaptivity.

- Out of sample prediction. Overfit in sample. Split into 3 parts.

  Short run versus long-run. Metrics of success.

- Computation. Algorithmic approach.

- Role for statistics? DGP/"true model".

  Challenges for inference. Adaptivity and model selection raise difficulties for the calculation of standard errors.

- Role for causal thinking? Recent work on treatment effects methods using ML (Athey, Imbens, Wager, Chernozhukov, Hansen...). Causation versus prediction.

- Role for economic models? Still a frontier area.

Lots of ideas are well-known in statistics, although the terminology differs between stats and ML. See Tibhirani's glossary: https://statweb.stanford.edu/ tibs/stat315a/glossary.pdf

Examples in Economics and outside

Outside econ: many successes. Ex: chess (new algorithms and new computers), pattern and speech recognition, text analysis (latent variable models and "denoising")...

Inside econ: not clear (yet?).

- Demand analysis: many attributes, many products.

- Dynamic games: large state space. Firm oligopoly (Benkard, Pakes, Weintraub,...).

- Network models (Graham, Jackson, Mele,...).

- New use of text data (Gentzkow, Shapiro and others).

- Econometric literature on Lasso (V. Chernozhukov and his team), but few applications so far.
- Related: model averaging. Read Sala i Martin (1997).
- Record linkage for matching data sets.
- Estimating average treatment effects under selection on observables (Athey, Imbens).

Outline of the course and references

- Regression with a large number of regressors.
- Penalized regression and the Lasso.
- Tree methods.
- Neural networks.
- Unobserved Heterogeneity.

References:

- Hastie, Tibshirani and Friedman: Elements of Statistical Learning.
- Hastie, Tibshirani Wainwright on sparsity and the Lasso.
- Murphy: Machine Learning, a probabilistic perspective.
- James, Witten, Hastie and Tibshirani: An Introduction to Statistical Learning with Applications in R.
- Many articles in econometrics, statistics and machine learning/computer science journals.

Chapter 1: Regression with a large number of regressors

1. **OLS with large p**

2. **Series regression**

OLS with large p

Model and examples

Linear model.

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + u_i.$$

We will assume that $\mathbb{E}(u_i \,|\, x_{i1}, ..., x_{iJ}) = 0$, and sometimes even the stronger assumption that $u_i \,|\, x_{i1}, ..., x_{iJ} \sim \mathcal{N}(0, \sigma^2)$. We will mostly abstract from issues of dependence between observations.

OLS.

$$(\widehat{\beta}_1, ..., \widehat{\beta}_J) = \underset{(b_1,...,b_J)}{\text{argmin}} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2.$$

Matrix form:

$$\widehat{\beta} = \underset{b}{\text{argmin}} \, \|y - X\beta\|^2 .$$

$$\widehat{\beta} = (X'X)^{-1} X'y.$$

Note that $X'X$ is singular when $p > n$.

Examples. (1a) Large number of regressors. Adding regressors may be helpful to make a conditional independence assumption more plausible.

(1b) Large number of instruments. Adding IVs in the first stage is tempting so as to increase the R^2. However there is a risk of overfitting the first stage. 2SLS is then biased toward OLS (Angrist and Krueger, 1991, Bound, Baker and Jaeger, 1995 JASA).

(2) Fixed-effects. Individuals, firms, cities, markets, products...

Incidental parameter problem. The number of observations may be small inside each cell.

Prevalence of unbalanced structures in empirical settings: the "long tail" phenomenon. It is appealing to use information from cells with many observations to learn about cells with few observations.

Let:

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + u_i,$$

where x_{ij} is binary and $x_{ij}x_{i\ell} = 0$ for all $j \neq \ell$.

Since the covariates are orthogonal, the OLS estimator is:

$$\widehat{\beta}_j = \overline{y}_j \text{ for all } j.$$

Hence:

$$\widehat{\beta}_j = \beta_j + \overline{u}_j.$$

If u_i are iid with zero mean and variance σ^2:

$$\text{Var}(\widehat{\beta}_j) = \frac{\sigma^2}{N_j},$$

where $N_j = \sum_{i=1}^{N} x_{ij}$ is the size of the j-cell.

Precision is driven by the number of observations in every cell.

(3) Nonparametric regression.

$$y_i = f(x_i) + u_i.$$

Here x_i has a small dimension. For example: x_i is scalar.

Approximate $f(x)$ in a basis of functions:

$$f(x) \approx \sum_{k=0}^{K} a_k \phi_k(x),$$

where $\phi_k(x)$ may be (1) ordinary polynomials, (2) Chebyshev, Legendre, Hermite... polynomials, (3) splines...

Here we need to let K grow to approximate f well.

Standard asymptotic analysis. Fixed p, N tends to infinity. If observations are i.i.d. then under standard conditions the CLT gives:

$$\sqrt{N}(\widehat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \mathbb{E}[x_i x_i']^{-1}\mathbb{E}[u_i^2 x_i x_i']\mathbb{E}[x_i x_i']^{-1}),$$

where x_i is $p \times 1$.

The White (1980) variance formula is:

$$(X'X)^{-1}X'\widehat{\Omega}X(X'X)^{-1},$$

where $\widehat{\Omega}$ is diagonal with elements $\widehat{u}_i = (y_i - x_i'\widehat{\beta})^2$.

Notice the presence of $(X'X)^{-1}$ in this expression.

Key point: this asymptotic expression is based on a thought experiment (with N tending to infinity). Is the resulting formula useful to capture the finite-sample variability of $\widehat{\beta}$?

This can be checked in specific DGPs using Monte Carlo simulations with different values of N and p.

High-dimensional asymptotic. When is p large?

Asymptotically, $p = p_N$.

This means that the population is changing with the sample size! In fact we are now considering a sequence of populations, indexed by N.

This approach captures finite sample performance better in cases where p is not negligibly small relative to N.

Similar devices have been used in the weak instruments literature (Staiger and Stock, 1997), panel data (Hahn and Newey, 2004, Arellano and Hahn, 2007), and the analysis of network data (Graham, 2017, and a large literature in statistics).

Formally, in this approach the rate of growth $p_N = p(N)$ is key: the faster p_N grows with N the more difficult the estimation problem is.

What a given rate p_N means for any given application is sometimes unclear. In practice we just have one N and one p. However, the high-dimensional asymptotic setup can motivate new estimators that tend to work better in applications. It is also very useful to construct confidence intervals that have a good coverage.

Behavior as p increases

Under normality.

$$u \,|\, X \sim \mathcal{N}(0, \sigma^2 I),$$

where the identity matrix I has dimension p.

Then:

$$\widehat{\beta} \,|\, X \sim \mathcal{N}(\beta, \sigma^2 (X'X)^{-1}).$$

Can be used to form exact confidence intervals.

Key question: is $X'X$ exactly or approximately collinear?

Checking collinearity on the computer. (1) Compute the rank of $X'X$ (numerically), (2) compute the eigenvalues of $X'X$.

Exact collinearity is very rare (but it happens! A classical example is the indeterminacy of age, time, and cohort effects in a linear regression). However: approximate multicollinearity is very frequent in applications.

Question: if X_1 and X_2 have a correlation of .99 should we still treat them as not collinear?

Important remark: (approximate) multicollinearity becomes more prevalent as the number of covariates increases.

Singular value decomposition. $X = UDV'$. U, V have orthogonal columns, D is diagonal, with dimensions equal to the rank of X.

We can order the diagonal elements of D in decreasing order. The key is the minimum element. Indeed:

$$(X'X)^{-1} = VD^{-2}V'.$$

Key issue: very often, when p is moderately large (say 20, 30) the minimum element of D is close to zero.

Remark: the SVD is a very useful tool when working with matrices more generally.

Huber's result. The rate of p_N as a function of N is key to ensure good performance of OLS.

In a classic paper, Huber (1973) shows that consistency and asymptotic normality of $\widehat{\beta}$ require p/N tending to zero.

Recent theory: Cattaneo, Jansson, Newey. Recently, Cattaneo, Jansson and Newey (2018) focus on one coefficient of β, say β_k, and show that consistency and asymptotic normality can still be achieved for such a one-dimensional parameter even when p/N tends to a non-zero constant.

Series regression

Approaches to nonparametric regression

Model. Let $f_0(x) = \mathbb{E}(y_i \,|\, x_i = x)$. We want to fit the model:

$$y_i = f_0(x_i) + u_i,$$

where $f_0(x) \approx \sum_{k=1}^{K} \beta_k \phi_k(x)$.

Kernel. Nadaraya-Watson estimator.

Local linear and local polynomial estimators.

These estimators achieve optimal rates in particular classes of functions f_0.

Series. OLS estimator: regress y_i on $\phi_0(x_i), ..., \phi_K(x_i)$.

The series estimator is computed given K. It is of course important to set a "good" value of K (see below).

We also need to choose a basis of functions $\phi_k(x)$. These can be standard polynomials, orthogonal polynomials, splines, wavelets... Different families have different approximation properties.

Series methods are not adaptive: the family ϕ_k is given, not estimated.

There exist penalized versions of series estimators, see next chapter.

Adaptive methods. In adaptive methods we estimate the basis functions $\phi_0, ..., \phi_K$.

Hence a model selection aspect to adaptive methods: what are "good" basis functions given my sample?

Example 1: regression tree.

Example 2: neural network.

Some theory on series

Read Newey (1997, Journal of Econometrics).

Newey establishes rates of convergence and asymptotic normality for series estimators.

An important result of the theory is a characterization of the ℓ^2 convergence rate:

$$\left\|\sum_{k=0}^{K} \widehat{a}_k \phi_K - f_0\right\|_2^2,$$

where the coefficients $\widehat{a}_0$, ..., $\widehat{a}_K$ are OLS estimates when regressing y_i on $\phi_0(x_i)$, ..., $\phi_K(x_i)$.

Assumptions. Let $\phi = (\phi_1, ..., \phi_K)'$. Note that ϕ depends on K.

Assumption 1: iid observations, $\mathrm{Var}(y \,|\, x)$ bounded.

Assumption 2: the smallest eigenvalue of $\mathbb{E}[\phi(x_i)\phi(x_i)']$ is $\geq c > 0$ for all K, $\sup_x \|\phi(x)\| \leq \zeta_K$, and $\zeta_K^2 K/N$ tends to zero.

Assumption 3: There are β_K and α such that $\sup_x |f_0(x) - \phi(x)'\beta_K| = O(K^{-\alpha})$ as $K \to \infty$.

This last assumption is about the smoothness of f_0.

Rates of convergence. Theorem (Newey): under Assumption 1-3:

$$\int \left(f_0(x) - \sum_{k=0}^{K} \widehat{a}_k \phi_k(x) \right)^2 dF(x) = O_p(K/N + K^{-2\alpha})$$

and:

$$\sup_x \left| f_0(x) - \sum_{k=0}^{K} \widehat{a}_k \phi_k(x) \right| = O_p(\zeta_K[\sqrt{K/N} + K^{-\alpha}]).$$

Intuition of the proof. Square norm, in case u_i are i.i.d. normal $(0, \sigma^2)$. For conciseness let us denote $\beta = \beta_K$ as in Assumption 3.

We are going to show a slightly different result as in part 1 of Newey's theorem, involving $\frac{1}{N}\sum_{i=1}^N \left(f_0(x_i) - \sum_{k=0}^K \widehat{a}_k \phi_k(x_i) \right)^2$ as opposed to $\int \left(f(x) - \sum_{k=0}^K \widehat{\beta}_k \phi_k(x) \right)^2 dF(x)$.

Applying the triangle inequality we have:

$$\begin{aligned} ||\phi\widehat{a} - f_0||_2^2 &\leq 2\,||\phi\beta - f_0||_2^2 + 2\,||\phi\widehat{a} - \phi\beta||_2^2 \\ &= O(K^{-2\alpha}) + \frac{2}{N}(\widehat{a} - \beta)'\phi\phi'(\widehat{a} - \beta), \end{aligned}$$

where the first term can be interpreted as a squared bias term, and the second term is a variance term. With some abuse of notation, here we have denoted as ϕ the $N \times (K+1)$ matrix with i-th row $(\phi_0(x_i), ..., \phi_K(x_i))$.

Let $\widetilde{a} = (\phi'\phi)^{-1}\phi'(\phi\beta + u)$. We have, conditioning on $x_1, ..., x_N$:

$$\begin{aligned} &\mathbb{E}\left[(\widetilde{a} - \beta)'\phi\phi'(\widetilde{a} - \beta)\right] = \mathbb{E}\left[u'\phi(\phi'\phi)^{-1}\phi' u\right] \\ &= \mathbb{E}\left[\mathrm{Tr}\left(\phi(\phi'\phi)^{-1}\phi' uu'\right)\right] = \sigma^2\,\mathrm{Tr}\left(\phi(\phi'\phi)^{-1}\phi'\right) \\ &= \sigma^2 K. \end{aligned}$$

The result then follows, provided that the difference:

$$(\widehat{a} - \beta)'\phi\phi'(\widehat{a} - \beta) - (\widetilde{a} - \beta)'\phi\phi'(\widetilde{a} - \beta)$$

is of smaller order (we can check it is).

Note: to show the result formally we need to check that $\phi'\phi$ is invertible with probability approaching one in large samples. This can be done using a trimming strategy (see

Newey, 1997, proof of Theorem 1).

Bias-variance trade-off. The square bias is proportional to $K^{-2\alpha}$ for some α that is larger, the "smoother" the true function f_0 is. The smoothness of a function measures how many terms ϕ_k are needed to approximate the function well.

The variance is proportional to K/N. This makes intuitive sense: we are estimating K parameters with N data points. Note that in the low-dimensional case (standard OLS theory) the variance is proportional to $1/N$ since K is a constant in that case.

Bias-variance trade-off. Bias(K) decreasing, Variance(K) increasing.

Choice of K. The bias-variance trade-off motivates selecting K "not too large, not too small". The theory provides some guidance on K: if the ℓ^2 rate is the desired objective, it should be selected to be proportional to $N^{1/(2\alpha+1)}$, so as to balance squared bias and variance.

However, the proportionality constant is unknown (and the "smoothness constant" α is also typically unknown). Selecting K in practice remains a challenging problem.

A popular approach is to rely on cross-validation.

Implementation of CV for series.

Average derivatives, asymptotic distribution. Newey (1997) also shows $\sqrt{N}$-normality asymptotic normality for functionals of f_0 such as average derivatives, under stronger assumptions. In particular he requires that $\sqrt{N}K^{-\alpha}$ tends to zero asymptotically.

He specializes these results for power series (i.e., ordinary polynomials) and regression splines.

Chapter 2: The Lasso and relatives

1. **Subset selection**
2. **The Lasso**
3. **Other penalization schemes**
4. **Matrix methods**

Subset selection

ℓ^0 penalty

Consider the problem:

$$\min_{b_1,\ldots,b_p} \sum_{i=1}^{N}(y_i - \sum_{j=1}^{p} b_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \mathbf{1}\{b_j \neq 0\}.$$

The ℓ^0 penalty differs from the ℓ^1 penalty that we will be using in Lasso since:

$$\|b\|_0 = \sum_{j=1}^{p} \mathbf{1}\{b_j \neq 0\}$$

only takes into account the number of non-zero b's ("extensive margin"), and discards the magnitudes of the coefficients ("intensive margin").

Link to information criteria (Akaike). The objective function is closely related to the AIC (Akaike) information criterion:

$$\min_{b_1,\ldots,b_p} \sum_{i=1}^{N}(y_i - \sum_{j=1}^{p} b_j x_{ij})^2 + 2\sigma^2 \sum_{j=1}^{p} \mathbf{1}\{b_j \neq 0\},$$

where $\sum_{j=1}^{p} \mathbf{1}\{b_j \neq 0\}$ is the number of parameters that we wish to estimate.

Here the criterion is the value of the minimum.

In practice, computing this criterion requires setting a value for σ^2, as usual when setting a penalty value.

The AIC is widely used for model selection. It can be used for more general objective functions such as likelihood, generalized method of moments (GMM), ...

The Bayesian information criterion (BIC) is a related, influential information criterion. It is justified differently.

Computation. Minimizing:

$$\|y - Xb\|_2^2 + \lambda \|b\|_0$$

is numerically hard.

This is because, unlike Lasso, the objective function is highly non-convex.

In non-convex settings, descent algorithms are often trapped in local optima. Starting the algorithm from several initial parameter values helps, but in high-dimensional settings the number of possible starting values is very large, and it is hard to assess whether we have reached the global optimum.

A possibility here is to fix the number of non-zero coefficients $\|b\|_0 = s$ and to minimize:

$$\|y - Xb\|_2^2, \quad \text{s.t. } \|b\|_0 \leq s.$$

This is called subset selection.

In subset selection, we compare the R^2 coefficients in any possible regressions with s covariates. The set of covariates with the highest R^2 is the one we select.

Remark: choosing s is analogous to choosing K in series.

Problem: subset selection is a combinatorial numerical problem (even too hard for “big computers”). The Lasso approximates the solution to this problem, yet it is much simpler computationally since its objective function is convex.

Stepwise methods. There exist some alternatives to exhaustive, combinatorial search. Heuristic methods include stepwise forward regression, for example (see chapter 3.3 in Hastie, Tibshirani and Friedman’s book). They are rarely used in econ.

Theory

A sparse Data Generating Process. Consider a model where the true parameters β_{j0} are such that at most s of them are zero, and all other β_{j0}’s are zero. The model is given by:

$$y_i = \sum_{j=1}^{p} \beta_{j0} x_{ij} + u_i,$$

where u_i is independent of $x_{i1}, ..., x_{ip}$ and normal with mean zero and variance one.

Case with $s = 1$ regressors to select. It is useful to consider a simple example where one wishes to select one covariate out of p, so $s = 1$.

In this case subset selection amounts to selecting the covariate x_{ij} that has the highest correlation with y_i.

If we assume that the error term u_i is $\mathcal{N}(0, \sigma^2)$ then it is possible to show that subset selection will select the "right" covariate, as N tends to infinity.

A large number of machine learning methods are based on such model selection ideas.

Consistency under normality. So see why this works, suppose in addition that covariates are standardized: $\sum_{i=1}^N x_{ij} = 0$ and $\sum_{i=1}^N x_{ij}^2 = 1$ for all j.

It is very common (and recommended) when applying high-dimensional regression methods to standardize your covariates in this way.

In this case subset selection is going to select:

$$\widehat{j} = j \Leftrightarrow \left| \sum_{i=1}^N y_i x_{ij} \right| \geq \left| \sum_{i=1}^N y_i x_{i\ell} \right| \qquad \text{for all } \ell \neq j.$$

Suppose that the true DGP is:

$$y_i = \beta_{j^*} x_{ij^*} + u_i.$$

Does $\widehat{j}$ tend to j^* asymptotically?

We have (conditional on x's):

$$\begin{aligned}
&\Pr(\widehat{j} \neq j^*) = 1 - \Pr(\widehat{j} = j^*) \\
&= 1 - \Pr\left(\left| \sum_{i=1}^N y_i x_{ij^*} \right| \geq \left| \sum_{i=1}^N y_i x_{i\ell} \right| \qquad \text{for all } \ell \neq j^* \right) \\
&= 1 - \Pr\left(\left| \beta_{j^*} + \sum_{i=1}^N u_i x_{ij^*} \right| \geq \left| \beta_{j^*} \sum_{i=1}^N x_{ij^*} x_{i\ell} + \sum_{i=1}^N u_i x_{i\ell} \right| \qquad \text{for all } \ell \neq j^* \right).
\end{aligned}$$

This is a probability of misclassification.

We need that the $x_{i\ell}$'s are not "too correlated". For example let us assume that:

$$\sup_{\ell \neq j^*} |\sum_{i=1}^N x_{ij^*} x_{i\ell}| \leq c < 1.$$

Then we can bound the misclassification probability by:

$$1 - \Pr\left(\sup_{\ell=1,\ldots,p}\left|\sum_{i=1}^{N} u_i x_{i\ell}\right| < (1-c)\left|\beta_{j^*}\right|\right).$$

This probability tends to zero under mild conditions on the relative rates of N and p. To see this, note that by the union bound:

$$\Pr\left(\sup_{\ell=1,\ldots,p}\left|\sum_{i=1}^{N} u_i x_{i\ell}\right| > (1-c)\left|\beta_{j^*}\right|\right) = \Pr\left(\exists \ell\ :\ \left|\sum_{i=1}^{N} u_i x_{i\ell}\right| > (1-c)\left|\beta_{j^*}\right|\right)$$
$$\leq p \sup_{\ell=1,\ldots,p} \Pr\left(\left|\sum_{i=1}^{N} u_i x_{i\ell}\right| > (1-c)\left|\beta_{j^*}\right|\right),$$

which, since $\sum_{i=1}^{N} u_i x_{i\ell}$ is a normal (0,1/N), tends to zero provided $p\Phi(N(1-c)\beta_{j^*})$ tends to zero.

This last condition holds whenever p/N^k tends to zero for some $k > 0$, for example. So it allows for $p >> N$.

Relaxing normality. We only used normality of u_i in the last step of the argument.

When u_i is not normal, one can use large deviations analysis to obtain bounds on:

$$\Pr\left(\left|\sum_{i=1}^{N} u_i x_{i\ell}\right| > (1-c)\left|\beta_{j^*}\right|\right).$$

Standard tools are exponential inequalities: Bernstein, Hoeffding.

The Lasso

Main ideas

The Lasso penalty. Replace the least squares minimization by:

$$\min_{b_1,\ldots,b_p} \sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{p} b_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p} |b_j|.$$

Read Tibshirani (1996), who originally proposed the Lasso.

In matrix form:

$$\min_b \|y - Xb\|_2^2 + \lambda\|b\|_1.$$

This is OLS + ℓ^1 penalty.

When $\lambda = 0$, the Lasso solution coincides with OLS.

When λ tends to infinity all $\widehat{\beta}_j$ tend to zero.

Norms in $\mathbb{R}^p$. ℓ^2, ℓ^1, ℓ_∞. Some inequalities.

Why sparse solutions? It is useful to consider the dual representation of the Lasso problem:

$$\min_b \|y - Xb\|_2^2, \quad \text{s.t. } \|b\|_1 \leq C.$$

There is a one-to-one mapping between C and λ. λ may be interpreted as the Lagrange multiplier associated with the constraint $\|b\|_1 \leq C$.

This representation gives insight about why the Lasso solutions tend to have a lot of zeros, i.e. to be "sparse".

The level curves of the OLS objective are ellipsoids. The constraints set is diamond-shaped. Like in standard utility maximization theory, corner solutions are likely.

In two dimensions (b_1, b_2), corner solutions correspond to either $b_1 = 0$ or $b_2 = 0$.

An example: estimating network links from panel data

Panel data model. Read Manresa (2013):

$$y_{it} = w_{it}'\delta + \sum_{j=1}^{p} \gamma_{ij} x_{jt} + \alpha_i + u_{it}.$$

Here y_{it} may be some measure of productivity of firm i at time t, x_{jt} is the amount of R& D innovation that firm j does in period t, and w_{it} are other determinants of productivity. α_i is a firm-specific fixed effect.

Panel Lasso estimator. Manresa estimates this model by minimizing:

$$\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - w_{it}'\delta - \sum_{j=1}^{p} \gamma_{ij} x_{jt} - \alpha_i\right)^2 + \lambda \sum_{j \neq i} |\gamma_{ij}|.$$

For simplicity assume that δ is known.

We can interpret this estimator as a collection of N Lasso estimators, each of which minimizes:

$$\sum_{t=1}^{T}\left(y_{it} - w_{it}'\delta - \sum_{j=1}^{p} \gamma_{ij} x_{jt} - \alpha_i\right)^2 + \lambda \sum_{j \neq i} |\gamma_{ij}|.$$

The objective function is convex and can be minimized using cyclic coordinate descent (see next chapter).

Application: R&D spillovers. The panel Lasso estimator will enforce that, for each firm j, most of the $\widehat{\gamma}_{ij}$ are zero.

Hence, each firm is responsive to at most a few R& D investments from other firms.

Sparsity is plausible in this setting, since it is likely that forming a link with another firm (which would allow firm i to benefit from j's research) is costly.

Other penalization schemes

Ridge regression

In Ridge regression, we use as a penalty the sum of the squares $\sum_{j=1}^{p} b_j^2$.

The Ridge objective.

$$\min_{b_1,\ldots,b_p} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} b_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} b_j^2.$$

This is a convex function. Strictly convex when $\lambda > 0$.

Dual formulation. In this case the dual is:

$$\min_{b_1,\ldots,b_p} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} b_j x_{ij} \right)^2 , \quad \text{s.t.} \sum_{j=1}^{p} b_j^2 \leq C.$$

The level curves of the OLS objective are ellipsoids, and the constraint set is a ball.

Hence, unlike Lasso Ridge regression does not produce sparse solutions in the sense that typically all Ridge estimates will be non-zero.

Orthogonal covariates. When regressors are orthogonal to each other the Ridge regression estimator has a closed-form solution:

$$\widehat{\beta}_j^{Ridge} = \frac{\widehat{\beta}_j^{OLS}}{1+\lambda}, \quad j = 1, ..., p.$$

Matrix formulation and solution. More generally, we can write Ridge as minimizing:

$$||y - Xb||^2 + \lambda ||b||^2.$$

The solution is:

$$\widehat{\beta}^{Ridge} = (X'X + \lambda I)^{-1} X'y,$$

where I is $p \times p$.

Notice that, even if $X'X$ is singular, $X'X + \lambda I$ is non-singular.

A SVD formulation. Let $X = UDV'$ be the SVD of X.

Here we write the "complete' SVD, so U is $N \times N$, D is $N \times p$ and V is $p \times p$.

We have:

$$\begin{aligned}(X'X + \lambda I)^{-1} X'y &= (VD'DV' + \lambda I)^{-1} VDU'y \\ &= V(D'D + \lambda I)^{-1} DU'y.\end{aligned}$$

Contrast this with the OLS estimator:

$$\begin{aligned}(X'X)^{-1} X'y &= (VD'DV')^{-1} VDU'y \\ &= V(D'D)^{-1} DU'y,\end{aligned}$$

which requires $D'D$ to be non-singular.

There is an interesting connection between Ridge regression and principal components (see Section 3.4 in the book by Hastie, Tibshirani and Friedman).

Choosing λ. As in Lasso, λ drives the bias/variance trade-off: small λ means high variance, high λ means high bias.

A popular approach is to select λ using cross-validation.

Variants of the Lasso.

Post Lasso. In practice one can proceed in two steps: (1) recover the relevant regressors using the Lasso, (2) run OLS on the set of relevant regressors.

This approach is referred to as "post-Lasso".

The hope with the post-Lasso estimator is that we select a suitable set of regressors, yet avoid "shrinking" the estimates too much towards zero.

Post-Lasso aims at decoupling the model selection and shrinkage aspects of Lasso.

Group Lasso. Suppose we have now two linear equations:

$$y_{i1} = \sum_{j=1}^{p} \beta_{j1} x_{ij1} + u_{i1}$$
$$y_{i2} = \sum_{j=1}^{p} \beta_{j2} x_{ij2} + u_{i2}.$$

An example is a panel data model with 2 periods.

In some applications one wishes to impose that (β_{j1}, β_{j2}) is either jointly zero, or jointly non-zero.

The group lasso penalty is then:

$$\sum_{j=1}^{p} \sqrt{b_{j1}^2 + b_{j2}^2}.$$

Note that this is also:

$$\sum_{j=1}^{p} \|b_j\|_2.$$

The objective functions is still convex, and the properties are similar to the ones of the Lasso.

Fused Lasso. Consider a time-series model:

$$y_t = \beta_t + u_t,$$

where u_t is an iid shock, and β_t is very persistent.

For example, we might want to smooth the evolution of GDP or inflation.

Using the fused Lasso, one can impose that β_t changes only a few times during the observation sample, by minimizing:

$$\sum_{t=1}^{T} (y_t - b_t)^2 + \lambda \sum_{t=2}^{T} |b_t - b_{t-1}|.$$

Here we are penalizing the increments of b_t.

This idea can be generalized to allow for time-varying coefficients in a regression, say $\beta_t x_t$ for example.

Elastic Net. With the Elastic net we combine the Lasso and Ridge penalities. Specifically, we compute:

$$\min_{b_1,\ldots,b_p} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} b_j x_{ij} \right)^2 + \lambda \left[(1-\alpha) \sum_{j=1}^{p} b_j^2 + \alpha \sum_{j=1}^{p} |b_j| \right],$$

for some $\alpha > 0$.

We obtain the Lasso or Ridge as special cases, depending on the value of α.

The hope relative to the Lasso is to capture better the presence of "small" non-zero coefficients.

Example: estimation of social interactions in De Paula, Rasul and Souza (2018).

Lasso in nonlinear models. The Lasso can be used in nonlinear problems too. For example, in logit we can compute:

$$\min_{b_1,\ldots,b_p} \sum_{i=1}^{N} y_i \ln \Lambda \left(\sum_{j=1}^{p} b_j x_{ij} \right) + (1-y_i) \ln \Lambda \left(- \sum_{j=1}^{p} b_j x_{ij} \right) + \lambda \sum_{j=1}^{p} |b_j|,$$

where $\Lambda(u) = \exp(u)/(1+\exp(u))$.

The same idea can be used in multinomial logit.

Example: high-dimensional demand model.

Likewise, we can Lasso-penalize Maximum Likelihood, minimum-distance, and GMM estimation problems.

Chapter 3: The Lasso: basic theory and implementation

1. **Convergence rates**
2. **Choice of λ**
3. **Computation of the Lasso**

Convergence rates

Read chapter 11 in Hastie, Tishirani and Wainwright.

Model and assumptions

A sparse DGP. Consider the model:

$$y = X\beta_0 + u,$$

where X is $N \times p$, and β_0 has $p - s$ elements that are zero.

s is the "sparsity level" of β_0.

We will assume that $u \,|\, X \sim \mathcal{N}(0, \sigma^2)$, but this can be generalized to allow for non-normality, heteroskedastivity, and some forms of dependence.

Restricted eigenvalue condition. Consider first the low-dimensional case. In this case, we impose that the minimum eigenvalue of $X'X$ is bounded away from zero; i.e.,

$$\frac{\nu' X' X \nu}{N \nu' \nu} \geq \gamma > 0, \quad \text{for all } \nu \neq 0.$$

This condition expresses the strict convexity of the least-squares objective.

In the high-dimensional case this condition does not hold in general. In fact the least-squares objective is often not strictly convex (think of the case $p > N$).

However, we can still have:

$$\frac{\nu' X' X \nu}{N \nu' \nu} \geq \gamma > 0, \quad \text{for all } \nu \neq 0 \in \mathcal{C},$$

where $\mathcal{C}$ is a suitable set.

Ideally, we would like to set $\mathcal{C}$ in such a way that only the ν vectors that have the same non-zero set as β_0 are considered. Denoting as S the set of non-zero elements in β_0, and as S^c its complement in $\{1, ..., p\}$, one would only look at $\nu \neq 0$ such that $\nu_j = 0$ for all $j \in S^c$.

It turns out that this condition is not sufficiently informative, but it is enough to relax it a little and define:

$$\mathcal{C} = \{\nu \in \mathbb{R}^p \,:\, \|\nu_{S^c}\|_1 \leq 3\|\nu_S\|_1\},$$

where ν_S is the vector ν with zeros in place of the j where $\beta_{0j} = 0$, and ν_{S^c} is the vector ν with zeros in place of the j where $\beta_{0j} \neq 0$.

We will assume that the restricted eigenvalue condition holds with this set $\mathcal{C}$.

Convergence rates

Here we characterize the ℓ^2 convergence rate of the Lasso estimator, which minimizes:

$$\frac{1}{2N}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1.$$

Note the factor $1/2N$, which we introduce here to be close to the notation of chapter 11 in Hastie, Tishirani and Wainwright.

A basic result. Theorem 11.1 in Hastie, Tishirani and Wainwright provides a consistency result for Lasso. It only requires the above restricted eigenvalue condition, and the assumptions that λ is chosen such that:

$$\lambda \geq \frac{2}{N}\|X'u\|_\infty.$$

The theorem says that:

$$\|\widehat{\beta}^{\text{Lasso}} - \beta_0\|_2 \leq C\sqrt{s}\lambda,$$

where $C > 0$ is a constant (it can be chosen as $C = 3/\gamma$).

Note that, by choosing $\lambda = \frac{2}{N}\|X'u\|_\infty$ we obtain:

$$\|\widehat{\beta}^{\text{Lasso}} - \beta_0\|_2 \leq C\sqrt{s}\frac{2}{N}\|X'u\|_\infty.$$

Proof. Let $\nu = \widehat{\beta}^{\text{Lasso}} - \beta_0$.

We have:

$$\frac{1}{2N}\|y - X\widehat{\beta}\|_2^2 + \lambda\|\widehat{\beta}\|_1 \leq \frac{1}{2N}\|y - X\beta_0\|_2^2 + \lambda\|\beta_0\|_1.$$

This can be written:

$$\frac{1}{2N}\nu' X' X \nu + \lambda \|\widehat{\beta}\|_1 \leq \frac{1}{N} u' X \nu + \lambda \|\beta_0\|_1.$$

We have (using the Holder inequality):

$$\frac{1}{2N}\nu' X' X \nu + \lambda \|\widehat{\beta}\|_1 \leq \frac{1}{N} \|X' u\|_\infty \|\nu\|_1 + \lambda \|\beta_0\|_1.$$

Suppose that $\lambda \geq \frac{2}{N}\|X'u\|_\infty$. Then:

$$\frac{1}{2N}\nu' X' X \nu + \lambda \|\widehat{\beta}\|_1 \leq \frac{\lambda}{2} \|\nu\|_1 + \lambda \|\beta_0\|_1.$$

Using the triangle inequality we have:

$$\frac{1}{2N}\nu' X' X \nu \leq \frac{\lambda}{2} \|\nu\|_1 + \lambda (\|\nu_S\|_1 - \|\nu_{S^c}\|_1).$$

First, we see that since:

$$\|\nu_{S^c}\|_1 \leq \|\nu_S\|_1 + \frac{1}{2}\|\nu\|_1$$

we have:

$$\|\nu_{S^c}\|_1 \leq 3\|\nu_S\|_1,$$

that is, ν belongs to $\mathcal{C}$.

Hence, applying the restricted eigenvalue condition we have:

$$\frac{\gamma}{2}\nu'\nu \leq \frac{\lambda}{2} \|\nu\|_1 + \lambda (\|\nu_S\|_1 - \|\nu_{S^c}\|_1).$$

Now, by the Cauchy-Schwarz inequality:

$$\|\nu_S\|_1 \leq \sqrt{s}\|\nu_S\|_2.$$

Hence, using that $\nu \in \mathcal{C}$ we have:

$$\frac{\gamma}{2}\nu'\nu \leq \frac{3\lambda}{2}\|\nu_S\|_1 \leq \frac{3\lambda}{2}\sqrt{s}\|\nu_S\|_2.$$

Hence:

$$\|\nu_S\|_2 \leq \frac{3}{\gamma}\sqrt{s}\lambda,$$

which is what we wanted to show.

Normal errors. Under normal error, we have, using the fact that the columns of X are normalized (which is a convention we always use) and that $u|X$ is normal:

$$\frac{2}{N}\|X'u\|_\infty = O_p\left(\sqrt{\frac{\ln p}{N}}\right).$$

Hence:

$$\|\widehat{\beta}^{\text{Lasso}} - \beta_0\|_2 = O_p\left(\sqrt{s\frac{\ln p}{N}}\right).$$

We see that we loose a little relative to the rate of OLS in the regression of y on the columns of X that have non-zero coefficients, since the rate for that estimator is $\sqrt{\frac{s}{N}}$.

Support recovery?

Irrepresentability condition. The Lasso estimator provides consistent estimates of the coefficients under relatively mild conditions.

However, stronger conditions are needed in order to show that the Lasso estimates correctly which regressors belong to the regression and which ones don't. This type of results is called "support recovery".

The first condition is an irrepresentability condition:

$$\max_{j\in S^c} \|(X_S^T X_S)^{-1} X_S' x_j\|_1 \leq 1-\gamma,$$

for some $\gamma > 0$.

Note this is the condition we assumed to show support recovery of subset selection when $s = 1$.

This is commonly interpreted as a form of "near orthogonality".

β-min condition. Another condition for exact recovery is a beta-min condition. This condition imposes that, among the β_{0j} that are non-zero, none of them is "too close" to zero (in a sense that can be made precise, see Theorem 11.3 in Hastie, Tibshirani and Wainwright).

Such a condition is intuitively needed for exact recovery, since models with many small coefficients may cause the Lasso to make many mistakes.

Although such mistakes may not affect the ℓ^2 consistency of the estimator, they will typically affect the estimation of the set of relevant coefficients.

Is exact recovery likely in applications? Several authors have tried to check in simulations how likely the conditions for exact recovery are to be met. Hastie, Tishirani and Wainwright present such a simulation.

More generally, the “mistakes” in recovering the relevant covariates are an important motivation for the development of modern inference methods for the Lasso, and for the introduction of double Lasso and de-sparsified Lasso techniques that we will see in the next chapter.

Except in low dimensional cases, exact recovery of Lasso is likely to fail.

Choice of λ

Analytical formula

Theory. A first possibility to select λ is to use the formula for the theory:

$$\lambda = \frac{2}{N}\|X'u\|_\infty.$$

This formula guarantees a low ℓ^2 error, see the above theorem. However, note that different objectives (e.g., constructing confidence intervals for a component of β_0) would typically require a different λ.

In practice, u is not known, but in the $\mathcal{N}(0, \sigma^2)$ case taking:

$$\lambda = 2\sigma\sqrt{\tau\frac{\ln p}{N}},$$

for some $\tau > 2$, will have theoretical guarantees for the ℓ^2 error.

This is the formula given in the seminal paper by Bickel, Ritov and Tsybakov (2009).

An obvious question is how to estimate σ. One could fit a regression with a “large set” of regressors, and estimate σ as the RMSE. However, it may not be clear how to select this “large set” of regressors

Some authors have developed generalization of these formulas allowing for heteroskedasticity or dependence.

Removing σ: square-root lasso. Belloni, Chernozhukov and Wang (2011) propose a small deviation of the Lasso that avoids the need to estimate σ.

Their square-root Lasso method aims at minimizing:

$$\|y - X\beta\|_2 + \lambda\|\beta\|_1.$$

The objective function remains convex, as in Lasso. Computation can be based on conic programming.

Moreover, when $u|X \sim \mathcal{N}(0, \sigma^2)$, λ can be chosen by a formula that does not depend on σ.

However, this simplification crucially relies on homoskedasticity.

Cross-validation

CV for Lasso. An alternative to the formula-based approach is to select λ using CV.

Theoretical motivation. Chetverikov and Liao (2018) show some theory for CV to choose λ in Lasso. They derive convergence rates of the Lasso with CV-selected penalty. The rates are close to the optimal ones.

Computation

Coordinate descent

Lasso with one covariate. Consider the case where $p = 1$ and x is univariate.

In this case the solution of the Lasso is available in closed-form.

Indeed, we want to minimize:

$$\min_b \|y - bX\|^2 + \lambda|b|.$$

We can plot the objective function as a function of b.

The solution is:

$$\widehat{\beta}^{Lasso} = sign(\widehat{\beta}^{OLS})(|\widehat{\beta}^{OLS}| - \lambda)^+.$$

Hence $\widehat{\beta}^{Lasso}$ and $\widehat{\beta}^{OLS}$ have the same sign.

The magnitude of $|\widehat{\beta}^{Lasso}|$ is always smaller than the magnitude of $|\widehat{\beta}^{OLS}|$. We say that the Lasso "shrinks" the OLS estimate towards zero.

The Lasso objective is convex. Convexity of the Lasso objective function is obvious in the univariate case.

Note that the Lasso objective is not continuously differentiable: there is non-differentiability at zero.

In the multivariate case, the Lasso objective is convex as well. This is very useful, as there are no multiple local solutions, and we can reach the global minimum of the function.

Cyclic coordinate descent. The idea of cyclic coordinate descent is to update each b_j one at a time. Given all other parameters except b_j, the update of b_j is simply based on the objective:

$$\min_{b_j} \sum_{i=1}^{N} \left(y_i - b_j x_{ij} - \sum_{\ell=1}^{p} \mathbf{1}\{\ell \neq j\} b_\ell x_{i\ell} \right)^2 + \lambda |b_j|.$$

This is a univariate Lasso problem.

The solution is:

$$b_j = sign(\widehat{\beta}_j)(|\widehat{\beta}_j| - \lambda)^+,$$

where $\widehat{\beta}_j = \frac{1}{N} \sum_{i=1}^{N} x_{ij}(y_i - \sum_{\ell=1}^{p} \mathbf{1}\{\ell \neq j\} b_\ell x_{i\ell})$.

We implement this method iteratively, starting from initial values of $b_1, ..., b_p$ and looping over $j = 1, ..., p, 1, ..., p, ...$ until numerical convergence.

There are other popular methods to estimate the Lasso, such as the LARS (least angle regression) algorithm.

Least angle regression. LARS is a popular alternative to cyclic coordinate descent. In particular, it is useful in order to compute the full path of solutions, as a function of λ.

Lasso graphs are often reported in the literature. The two books by Hastie and coauthors provide illustrative applications.

Chapter 4: Inference on the Lasso

1. **Post-model selection inference**

2. **Double Lasso**

Post-model selection inference

Read Leeb and Potscher (2005).

A simple example

Model. Consider the following model:

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

where we abstract from the intercept for simplicity.

We are interested in β_1 only.

Pre-test estimator. Let $\widehat{\beta}_1^{long}, \widehat{\beta}_2^{long}$ denote the OLS estimator in the long regression of y_i on x_{i1}, x_{i2}.

Let $\widehat{\beta}_1^{short}$ denote the OLS estimator in the short regression of y_i on x_{i1}.

We study the properties of the following pre-test estimator: if $|\widehat{\beta}_2^{long}| > c_N/\sqrt{N}$ then report $\widehat{\beta}_1^{long}$. Otherwise report $\widehat{\beta}_1^{short}$.

Note the similarity with Lasso thresholding.

We choose c_N such that (1) $c_N \to \infty$, (2) $c_N/\sqrt{N} \to 0$. An example is the BIC criterion, where c_N is proportional to $\ln N$.

Lessons from standard asymptotics. Suppose β_2 is given, independent of the sample size.

Then $\Pr(|\widehat{\beta}_2^{long}| \leq c_N/\sqrt{N} \,|\, \beta_2 \neq 0)$ tends to zero, by (2).

Moreover, $\Pr(|\widehat{\beta}_2^{long}| > c_N/\sqrt{N} \,|\, \beta_2 = 0)$ tends to zero as well, by (1).

This shows that as N tends to infinity we select the "right" (long or short) model.

Hence the pre-test estimator seems to dominate the "long" OLS.

We generally refer to such results as "oracle" results.

Under oracle results, performing inference is easy since one can proceed as if the model was known.

However this logic is deceiving.

Problem with standard asymptotics. The problem with the previous argument is that it only works when β_2 is "very far away" from zero.

In finite samples, the distribution of the pre-test estimator is a mixture of two OLS estimators: based on the short and long regressions.

The standard asymptotic approach implies that the distribution is approximately normal. However the quality of the approximation may be poor in finite samples.

Indeed, although mistakes are rare, the cost associated with those mistakes (i.e., failing to include x_{i2} when it in fact affects the outcome) is severe.

The problem with pre-testing is pervasive in machine learning, since most ML methods involve some form of model selection.

Alternative asymptotic. Suppose that we now let $\beta_2 = \beta_{2N}$ depend on N. To fix ideas, let $\beta_{2N} = a/\sqrt{N}$ for a constant $a \neq 0$.

Then $\Pr(|\widehat{\beta}_2^{long}| \leq c_N/\sqrt{N} \,|\, \beta_2 = a/\sqrt{N},\, a \neq 0)$ tends to one as N tends to infinity.

Hence the pre-test estimator is always based on the short regression, even though $\beta_2 \neq 0$.

This alternative asymptotic regime highlights the problem with pre-test.

Alternatives to pre-test? In the case of the simple regression model, there is an obvious solution to the problem with pre-test: always report $\widehat{\beta}_1^{long}$.

This strategy may not be available in high-dimensional regression, however.

An alternative approach, which as we will see generalizes nicely to high-dimensional settings, is to compute a "double pre-test" estimator.

The idea is to report $\widehat{\beta}_1^{short}$ only if both $|\widehat{\beta}_2^{long}| < c_N/\sqrt{N}$ and $|\widehat{\gamma}| < d_N/sqrtN$, where $\widehat{\gamma}$ is the OLS estimator in the regression of x_{i1} and x_{i2}, and d_N satisfies similar conditions as c_N.

The insight is the fact that we do not want to miss out on covariates x_{i2} that are highly correlated with x_{i1}, because mistakes in that case would have severe consequences (i.e., a large omitted variable bias).

Examples of pre-testing issues

Pre-testing issues are prevalent in applications.

Hausman test. Hausman tests of exogeneity are very frequent in applied work. When the test does not reject exogeneity, it is common to report OLS instead of IV. Likewise, in panel data applications it is common to report random-effects estimators when we fail to reject that they differ from the fixed-effects estimators. Patrik Guggenberger has studied these issues.

F-test in IV. In IV regressions, it is common to test for weak instruments using a threshold of 10 for the F statistic, and then report IV when the test rejects that the instrument is irrelevant. Isaiah Andrews has studied this issue.

Use of information criteria for model selection. Information criteria (AIC, BIC, Mallows...) are widely used in applied work. One typically computes standard errors of estimators as if the true model was the one selected by the information criteria.

Double Lasso

Read Belloni, Chernozhukov and Hansen (2014, BCH).

Double Lasso in high-dimensional regression

Regression model and target parameter. Consider now the high-dimensional model:

$$y_i = \beta_1 x_{i1} + \sum_{j=2}^{p} \beta_j x_{ij} + u_i.$$

Suppose that we are interested in β_1, and interpret the other covariates as controls.

Suppose that the number of β_j's that are non-zero is "small" (e.g., fixed as the sample size tends to infinity).

Compute standard errors for Lasso "naively"? A possibility is to run Lasso. We know that under mild conditions Lasso is consistent.

To compute standard errors, it is tempting to report usual standard errors, treating the covariates selected by the Lasso as the "true" covariates in the model.

However, as pointed out by BCH doing this runs into a pre-testing issue.

Double Lasso estimator. BCH propose a double Lasso estimator:

(1) run Lasso of y_i on x_{i2},...,x_{ip}. Let $\widehat{S}_1$ denote the set of covariates recovered by the Lasso.

(2) run Lasso of x_{i1} on x_{i2},...,x_{ip}. Let $\widehat{S}_2$ denote the set of covariates recovered by the Lasso.

(3) run OLS of y_i on x_{i1} and all the covariates in $S_1 \cup S_2$.

Intuition: the low-dimensional case. To provide intuition on why this might work, consider again the simple model:

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

In this case, when we make the mistake of not including x_{i2} the omitted variable bias, which is $\beta_2 \gamma$, is of order $O(1/N)$, which is very small.

In contrast, with pre-test the omitted variable bias is of order $O(1/\sqrt{N})$.

The second regression (or Lasso) gives us a form of insurance against the presence of model selection mistakes.

Extra assumption: sparsity in the first stage. For the method to work we need the second Lasso to be well-behaved.

For this, we need to assume sparsity in the relationship:

$$x_{i1} = \sum_{j=2}^{p} \gamma_j x_{ij} + v_i.$$

Such a sparse relationship was not needed to establish consistency of the Lasso.

A related approach is the "desparsified Lasso" of Buhlman and Van de Geer. This method similarly relies on an extra assumption in the first stage, i.e. in the regression model of x_{i1} on x_{ij}.

Main result. The main result in BCH is that, as N tends to infinity, the distribution of the double-Lasso estimator has the standard asymptotic form.

Hence one can estimate standard errors using the White formula, for example.

Moreover, this asymptotic convergence holds uniformly in a large class of data generating processes. This supports the claim that the method is helpful to deal with the issues of pre-test, which suffers from lack of uniformity.

Behavior in simulations. BCH provide simulations that suggest that double Lasso can be approximately normally distributed, even in situations where the Lasso estimator has a multi-modal distribution.

A recent working paper by Wuthrich and Zhu (2019) provides simulations that highlight some problems with Lasso-based methods, including double Lasso, and show some superiority of OLS-based inference.

Double Machine Learning. Read Chernozhukov, Duflo and co-authors (2017, Econometrics Journal). In that paper, the authors generalize the double Lasso method of BCH to a number of other settings.

Their approach relies on two key features. The first is orthogonalization. In regression, orthogonalization leads to considering two Lasso regressions instead of one, following the logic of partitioned regression/Frisch Waugh.

The second key feature is sample splitting. This is especially attractive when the first stage is not a Lasso estimator but a complex ML estimator for which distribution theory is not available. As we will see in the next chapter, sample splitting is helpful to perform valid inference in complex problems.

Chapter 5: Partitioning methods

1. **Methods based on covariates: kernel, nearest neighbor and kmeans**

2. **Decision trees: CART**

3. **Bagging and random forests**

Methods based on covariates: kernel, nearest neighbor and kmeans

Nonlinear regression. We focus on a nonlinear regression model:

$$y_i = f(x_i) + u_i.$$

Unlike basis expansion methods, here we allow for models where the choice of basis is adaptive, i.e. where we estimate the basis.

Methods with X property. We will distinguish two cases: methods where we use only data on covariates to estimate the basis, and methods where we use data on both covariates and outcomes. The first type of methods is sometimes called "methods with the X property", as in the seminal book by Devroye, Gyorfi and Lugosi (1996).

We first start by reviewing non-adaptive methods.

Step functions. A simple possibility is to use a step function estimator. In the univariate case this amounts to partitioning the real line or a compact interval. In the multivariate case there is a curse of dimensionality, since cells may be empty.

Optimal steps? When choosing the number of cells we face the standard bias-variance trade-off. Cross-validation is an option to select the step size.

Nadaraya Watson and kernel methods. A related possibility is Nadaraya-Watson. With a kernel function κ and a bandwidth h we compute:

$$\widehat{f}(x) = \frac{\sum_{i=1}^{N} \kappa(\frac{x-x_i}{h}) y_i}{\sum_{i=1}^{N} \kappa(\frac{x-x_i}{h})}.$$

Properties have been studied extensively. The bias is $O(h^{-2})$ for standard kernels, the variance is $O(1/(Nh))$.

An issue with kernel is that the bandwidth does not depend on x. It is possible to allow for "adaptive bandwidths", but theory and implementation is harder.

Nearest neighbors. A popular adaptive method is K-nearest neighbors.

$$\widehat{f}(x) = \frac{\sum_{i=1}^{N} \mathbf{1}\{x_i \in B_x\} y_i}{\sum_{i=1}^{N} \mathbf{1}\{x_i \in B_x\}},$$

where B_x consists of the K points x_i closest to x according to some metric (example: in terms of Euclidean norm).

This is adaptive, since the width of B_x varies with how dense the data is around x.

This is a method with the X property.

For consistency we need two key conditions: K tends to infinity, and diameter(B_x) tends to zero.

Inference in nearest neighbor methods is challenging.

Close relative: nearest neighbor matching. See Abadie and Imbens's work on asymptotic properties.

K-means algorithm. Another method is to define $\widehat{f}(x)$ as a piecewise-constant estimator on a partition of $x_1, ..., x_N$.

To obtain a partition with K elements, the k-means algorithm is useful.

The k-means algorithm is obtained by:

$$\left(\widehat{h}, \widehat{k}_1, ..., \widehat{k}_N\right) = \underset{\left(\widetilde{h}, k_1, ..., k_N\right)}{\operatorname{argmin}} \sum_{i=1}^{N} \left\| x_i - \widetilde{h}(k_i) \right\|^2 .$$

The minimum is over all possible partitions of units in at most K groups. Computing a global minimum may be challenging, yet fast and stable heuristic algorithms have been developed, such as iterative descent, genetic algorithms or variable neighborhood search. Lloyd's algorithm is often considered to be a simple and reliable benchmark (see Steinley, 2006, and Bonhomme and Manresa, 2015, for algorithms and references).

The result of k-means is a Voronoi Tesselation.

Adaptivity of k-means. In general, the k-means objective function is of the order $K^{-2/d}$ where d is the dimension of x_i.

However, suppose $x_i = \varphi(\xi_i)$ where ξ_i is scalar and φ is Lipschitz-continuous. Then the k-means objective function is of the order K^{-2}.

K-means "adapts" to the low dimensionality of the data. This is very different from kernel methods, for example.

K-means partition of X. Given a k-means partition $\{\widehat{k}_i\}$, we compute:

$$\overline{y}(\widehat{k}_i),$$

the mean of y_i in group $\widehat{k}_i$.

To show consistency, note that letting $z_i = f_0(x_i)$:

$$\frac{1}{N}\sum_i \|\overline{y}(\widehat{k}_i) - f_0(x_i)\|^2 \leq \frac{2}{N}\sum_i \|\overline{z}(\widehat{k}_i) - z_i\|^2 + \frac{2}{N}\sum_i \|\overline{u}(\widehat{k}_i)\|^2.$$

The first term is the squared bias, bounded by $O_p(K^{-2/d})$ where d is the dimension of x_i.

Remark: the bias may be $>> 1/K^2$ since it is not clear how to partition z_i directly.

The second term is a variance. Since the method has the X property the variance is easy to control. For this suppose for example that $u_i \,|\, x_i$ is normal $0, \sigma^2$. In this case it can be shown that:

$$\frac{1}{N}\sum_i \|\overline{u}(\widehat{k}_i)\|^2 = O_p(K\sigma^2/N).$$

Hence the estimator is consistent provided K tends to infinity, and N/K tends to infinity.

When does using outcomes help? Only in some DGPs. The key idea here is still adaptivity. However, we may need to adapt to the structure of $Y \,|\, X$, as opposed to the structure of X.

As an example, consider regression model where: $\mathbb{E}(Y \,|\, X_1, X_2) = \mathbb{E}(Y \,|\, X_1)$. In this model, when X_1, X_2 have an absolutely continuous distribution, say, a method with the X property always faces a three-dimensional problem.

In contrast, it may be that a method with the Y, X property is able to learn that X_2 does not affect Y given X_1. In that case one hopes that the method could face only a two-dimensional problem.

Decision trees are a very simple set of methods with the Y, X property.

However, note right away the challenge: if we use Y in order to partition X, then performing inference might be harder.

A failed first attempt. What about computing:

$$\left(\widehat{h},\widehat{k}_1,...,\widehat{k}_N\right)=\underset{\left(\widetilde{h},k_1,...,k_N\right)}{\text{argmin}}\ \sum_{i=1}^{N}\left\|y_i-\widetilde{h}(k_i)\right\|^2.$$

This is a good idea in some panel data settings (Bonhomme, Lamadon and Manresa, 2017), but this is not useful in cross-sectional regression.

The problem is that the partition of units that we get may be very ill-behaved in terms of x_i.

In other words, the partition is constructed based on $f_0(x_i)$ and u_i. The noise drives the partition in a first-order way.

A tree will be essentially the same idea, but imposing that the partitions be well-behaved in the X space (e.g., family of rectangles). Limiting the complexity of the family of partitions will be enough to restore consistency, and possibly exploit adaptivity of the fact that outcomes are used in the partitioning.

Decision trees: CART

Objective function. CART (classification and regression trees) was proposed by Breiman et al in their classic book (1984). The idea of a tree is to minimize:

$$\sum_{i=1}^{N}\left\|y_i-\widetilde{h}(k_i)\right\|^2,$$

subject to the k_i forming a partition of the X space that is not "too complex".

The restriction will be that the partition is composed of rectangles in the X space.

Greedy algorithm. In CART we first estimate a "deep" tree. The idea is as follows:

-Go though all possible covariates,

-and go through all possible splits of the covariates. A split is a threshold c that gives rise to two sub-samples $x_{ik}\leq c$ and $x_{ik}>c$.

-Choose the covariate and split that gives the lowest value of the within-group sum-of-squares.

-Continue until each leaf of the tree has one element.

Trees have a recursive structure that is helpful for fast computation, and interpretable (example: used for medical decision making).

Note that this algorithm is greedy, i.e. non-optimal. Finding the optimal partition among all sets of rectangles is a combinatorially hard problem.

Deep versus shallow trees. A deep tree is likely to overfit. At the bottom of the tree, all leafs have one observation so there is massive overfit.

In contrast, a shallow tree may have large leaves but the bias may be large. This is the familiar trade-off.

Pruning and cross-validation. In CART we first estimate a deep tree and then go back to prune some of the leaves of the tree. The idea is to consider all potential pruning of the deep tree as a set of potential trees, and to select the tree that achieves a desired fit/complexity balance.

A key result in Breiman's book shows that exploring the family of trees is in fact computationally feasible. Typically the chosen tree is based on a penalty term that penalizes the number of leaves. That term is often chosen by cross-validation.

The whole procedure, including the CV step, is typically very fast.

A difficulty. Trees may be inefficient is some standard models. Consider for example an additive model of the form: $\mathbb{E}(y_i \,|\, x_{i1}, x_{i2}) = x_{i1} + x_{i2}$. In this case one will typically need lots of rectangles to approximate the conditional expectation well. An obvious solution would be to add $x_{i1} + x_{i2}$ as a covariate, but it may be hard to spot whether a particular interaction is suitable or not. Neural networks address this concern to some extent by working with more flexible bases.

Extensions of CART. A Bayesian extension of CART is BART. Other extensions involve considering extended sets of bases, beyond rectangles.

Bagging and random forests

Bagging estimators

The bagging principle. Most often, we use the bootstrap to compute standard errors and compute confidence intervals.

In bagging, the idea is to use the bootstrap for estimation.

Given a bootstrap sample y_i^s, x_i^s, $i = 1, ..., N$, we estimate a bootstrap estimator $\widehat{\theta}^{(s)}$.

Then we report $\frac{1}{S}\sum_{s=1}^{S}\widehat{\theta}^{(s)}$.

Bagging in linear models. Bagging does not help in linear models. An exception where bagging may be beneficial is to "smooth out" the impact of outlying observations.

Bagging in nonlinear models. Bagging may help in nonlinear models.

The insight is that bagging may reduce variance, by averaging.

For the variance reduction to be substantial, it is important that the bootstrap samples be not too correlated.

Ensuring weak correlation between bootstrap samples is the key idea behind random forests.

Bagging trees: random forests

Bagged trees. In a forest, we try to average the results of trees that are weakly correlated.

Usually, we combine two ideas to ensure weak correlation, both due to Breiman.

(1) We construct the trees on bootstrap samples of the original sample. Of course theses samples are not independent since they are based on the same underlying data.

(2) When going down the tree, when deciding which covariate to split, we select a set of M potential covariates along which the split can occur. This selection is at random. Then we proceed with the split as usual. We proceed in this way before every split.

(2) is the crucial step. Intuitively, the lower M the less correlated the trees are. Hence M is a key tuning parameter. When M is equal to the number of covariates, the result of bagging is the original deep tree.

Athey and Wager (2017) provide asymptotic theory for random forests. In the next chapter we will see some theory for partitioning methods.

Practical issues for random forests. In practice researchers often do not prune, and simply average the resulting deep trees.

The choice of M is key. A common rule is the number of covariates divided by three.

An example: estimating heterogeneous treatment effects. Read Athey and Wager (2017).

Chapter 6: Some theory on trees

1. **Rate of convergence for regression trees**

2. **Honest inference**

Rate of convergence for regression trees

Tree partitions versus unrestricted partitions. In CART we consider partitions that are formed of rectangles in the X space.

The number of unrestricted partitions of X into K groups is of the order of K^N.

For rectangles, the number of partitions depends on the dimension of X. In the scalar case the number is of the order of N^K. In the d-dimensional case it is of the order of N^{Kd}.

Model. We consider the model:

$$y_i = f_0(x_i) + u_i,$$

where $u_i \,|\, x_i$ is iid normal $(0, \sigma^2)$. We denote $z_i = f_0(x_i)$.

Estimator. We will analyze a tree partition that minimizes, for given K:

$$\sum_{i=1}^{N} \left\| y_i - \widetilde{h}(k_i) \right\|^2$$

subject to the partition $\{k_i\}$ consisting of rectangles in the X space.

There are two major differences with CART. (1) K is fixed ex ante here, and a single tree is grown given K. (2) We study the properties of the best tree according to the objective function (i.e., we are not using a greedy heuristic).

Consistency: first try. As in the case of k-means partition, we have:

$$\frac{1}{N}\sum_i \|\overline{y}(\widehat{k}_i) - f_0(x_i)\|^2 \leq \frac{2}{N}\sum_i \|\overline{z}(\widehat{k}_i) - z_i\|^2 + \frac{2}{N}\sum_i \|\overline{u}(\widehat{k}_i)\|^2.$$

However, here $\widehat{k}_i$ is dependent on the u_i. Hence, bounding the error is harder than in the case of k-means partition.

Infeasible partition. We consider a second, "infeasible" partition that minimizes:

$$\sum_{i=1}^{N}(z_i - \overline{z}(k_i))^2,$$

over the set of rectangles in the X space. We call this partition $\widetilde{k}_i$.

Note two key differences with the k-means case: (1) this partition is directly based on z_i, not on x_i, and (2) the minimization is only with respect to rectangles, not general partitions.

Bound. From:

$$\frac{1}{N}\sum_i \|y_i - \overline{y}(\widehat{k}_i)\|^2 \leq \frac{1}{N}\sum_i \|y_i - \overline{y}(\widetilde{k}_i)\|^2,$$

we obtain:

$$\frac{1}{N}\sum_i \|\overline{y}(\widehat{k}_i) - z_i\|^2 \leq \frac{1}{N}\sum_i \|\overline{y}(\widetilde{k}_i) - z_i\|^2 + \frac{2}{N}\sum_i u_i\left(\overline{y}(\widehat{k}_i) - \overline{y}(\widetilde{k}_i\right).$$

First term. We have, by the triangle inequality:

$$\frac{1}{N}\sum_i \|\overline{y}(\widetilde{k}_i) - z_i\|^2 \leq \frac{2}{N}\sum_i \|\overline{u}(\widetilde{k}_i)\|^2 + \frac{2}{N}\sum_i \|\overline{z}(\widetilde{k}_i) - z_i\|^2,$$

where:

$$\frac{1}{N}\sum_i \|\overline{u}(\widetilde{k}_i)\|^2 = O_p(K/N),$$

and the second term is a decreasing function of K.

Note that $\frac{1}{N}\sum_i \|\overline{u}(\widetilde{k}_i)\|^2 = O_p(K/N)$ comes from the fact that the u_i are independent of the $\widetilde{k}_i$ (although they are not independent of the $\widehat{k}_i$).

Second term. We have:

$$\frac{1}{N}\sum_i u_i\left(\overline{y}(\widehat{k}_i) - \overline{y}(\widetilde{k}_i\right) = \frac{1}{N}\sum_i u_i\left(\overline{z}(\widehat{k}_i) - \overline{z}(\widetilde{k}_i\right) + \frac{1}{N}\sum_i u_i\left(\overline{u}(\widehat{k}_i) - \overline{u}(\widetilde{k}_i\right).$$

Now, by the above:

$$\frac{1}{N}\sum_i u_i\left(\overline{u}(\widehat{k}_i) - \overline{u}(\widetilde{k}_i\right) = \frac{1}{N}\sum_i \overline{u}(\widehat{k}_i)^2 + O_p(K/N).$$

Union bound. To bound: $\frac{1}{N}\sum_i \overline{u}(\widehat{k}_i)^2$, we use the union bound. That is, for any $\epsilon > 0$:

$$\Pr\left(\frac{1}{N}\sum_i \overline{u}(\widehat{k}_i)^2 > \epsilon\right) \leq P_{KN}\Pr\left(\chi^2_K > \epsilon N\right),$$

where P_{KN} is the number of rectangle-partitions with K pieces.

Number of rectangle-partitions. When x_i is scalar and $K = 2$, there are N possible rectangle-partitions. With K groups there are of the order of N^K.

When x_i has dimension d there are of the order of Nd^K rectangle-partitions.

This is compared to K^N unrestricted partitions.

Second term (continued). Now:

$$\Pr\left(\chi^2_K > \epsilon N\right) \leq \exp(-c\epsilon N/K),$$

for some $c > 0$. Hence:

$$\Pr\left(\frac{1}{N}\sum_i \overline{u}(\widehat{k}_i)^2 > \epsilon\right) \leq \exp(K\ln(Nd) - c\epsilon N/K),$$

which tends to zero provided $K^2\ln(Nd)/N$ tends to zero.

First part of the second term. Moreover:

$$\frac{1}{N}\sum_i u_i\left(\overline{z}(\widehat{k}_i) - \overline{z}(\widetilde{k}_i)\right) = \frac{1}{N}\sum_i \overline{u}(\widehat{k}_i,\widetilde{k}_i)\left(\overline{z}(\widehat{k}_i) - \overline{z}(\widetilde{k}_i)\right),$$

where $\overline{u}(\widehat{k}_i,\widetilde{k}_i)$ is them mean of u_i in the intersection of the two partitions.

By Cauchy Schwarz:

$$\left(\frac{1}{N}\sum_i \overline{u}(\widehat{k}_i,\widetilde{k}_i)\left(\overline{z}(\widehat{k}_i) - \overline{z}(\widetilde{k}_i)\right)\right)^2 \leq \frac{1}{N}\sum_i \overline{u}(\widehat{k}_i,\widetilde{k}_i)^2\frac{1}{N}\sum_i\left(\overline{z}(\widehat{k}_i) - \overline{z}(\widetilde{k}_i)\right)^2$$
$$\leq \frac{2}{N}\sum_i \overline{u}(\widehat{k}_i,\widetilde{k}_i)^2\left[\frac{1}{N}\sum_i\left(z_i - \overline{z}(\widetilde{k}_i)\right)^2 + \frac{1}{N}\sum_i\left(z_i - \overline{z}(\widehat{k}_i)\right)^2\right].$$

Note that $\frac{1}{N}\sum_i\left(z_i - \overline{z}(\widehat{k}_i)\right)^2$ is what we want to bound.

We can verify that it is enough to bound:

$$\frac{1}{N}\sum_i \overline{u}(\widehat{k}_i, \widetilde{k}_i)^2.$$

To do this, we use the union bound once more:

$$\Pr\left(\frac{1}{N}\sum_i \overline{u}(\widehat{k}_i, \widetilde{k}_i)^2 > \epsilon\right) \leq P_{KN} \Pr\left(\chi^2_{K^2} > \epsilon N\right),$$

which tends to zero provided that $K^3 \ln(Nd)/N$ tends to zero.

Honest inference

The honesty principle. In a "honest" tree, we estimate the partition on a first random subsample of the data. We then estimate the leaf-specific means using only the rest of the sample.

Computing confidence intervals. To see why this is useful, consider the question of constructing a confidence interval for $f_0(x)$.

Let $k_i(x)$ denote the group membership indicator of the piece that contains x. We want to compute:

$$\Pr(|\overline{y}(k_i(x)) - f_0(x)| \leq a).$$

This is bounded by:

$$\Pr(|\overline{u}(k_i(x))| \leq a + |\overline{z}(k_i(x)) - f_0(x)|).$$

Suppose that the partition is sufficiently fine such that $|\overline{z}(k_i(x)) - f_0(x)|$ can be neglected. This requires choosing a large K, and is akin to "under-smoothing" in classical nonparametric statistics.

In that case we can approximate the probability by:

$$\Pr(|\overline{u}(k_i(x))| \leq a).$$

If the partition is constructed using the same units as when calculating $\overline{y}$ and $\overline{u}$, then this probability is very hard to characterize.

In contrast, in an “honnest” tree:

$$\overline{u}(k_i(x)) \sim \mathcal{N}\left(0, \frac{\sigma^2}{N_x}\right),$$

where N_x is the number of observations in the piece that contains x.

Hence:

$$\Pr(|\overline{y}(k_i(x)) - f_0(x)| \leq a) \approx \Phi\left(\sqrt{N_x}\frac{a}{\sigma}\right),$$

and an approximate 95 confidence interval is constructed as:

$$\left[\overline{y}(k_i(x)) - 1, .96\frac{\sigma}{\sqrt{N_x}}, \overline{y}(k_i(x)) + 1, .96\frac{\sigma}{\sqrt{N_x}}\right].$$

In practice we need to estimate σ, for example using:

$$\widehat{\sigma}^2 = \frac{1}{N}\sum_i \left(y_i - \overline{y}(\widehat{k}_i)\right)^2.$$

Modifying CART. It is also possible to modify the CART algorithm to incorporate honest splitting at each step. Read Athey and Imbens (2016).

Inference for random forests Honesty is also key to establish that random forests provide asymptotically valid inference. Read Athey and Wager (2017).

Chapter 7: Neural networks

1. **Neural network models**

2. **Computation: backpropagation and stochastic gradient descent**

3. **Approximation property and issues in regularization and inference**

Neural network models

Nonlinear regression: basis of functions. A popular approach to nonlinear regression is to consider a basis of functions, and consider the following model:

$$y_i = \sum_{k=0}^{K} a_k \phi_k(x_i) + u_i.$$

As we saw, example of functions ϕ_k are ordinary polynomials, orthogonal polynomials, splines, and wavelets.

Nonlinear regression: adaptive basis. Here we consider a modification of this approach, where we aim at estimating the basis function.

Formally, we consider a model of the form:

$$y_i = \sum_{k=0}^{K} a_k \phi_k(x_i; \theta_k) + u_i.$$

As an example, ϕ_k could be the density of a normal distribution, and θ_k could denote its mean and variance.

Other popular examples are neural networks, which we consider in this chapter, and trees, which we will focus on in the next chapter.

The sigmoid function. Neural networks make extensive use of link functions. A link function is a nonlinear function that is used to specify a basis.

A popular example is the sigmoid:

$$\sigma(v) = \frac{\exp(v)}{\exp(v) + 1}.$$

Note that σ is simply the cdf of the standard logistic. It is increasing from 0 to 1.

Single-layer neural networks. A simple neural network is as follows. We start with the vector x_i, which is $p \times 1$.

We form K linear combinations of them:

$$z_{ik} = \sum_{j=1}^{p} \gamma_{jk} x_{ij}, \quad k = 1, ..., K.$$

Finally, we combine the z_k as follows, and write:

$$y_i = \sum_{k=1}^{K} a_k \sigma(z_{ik}) + u_i.$$

Note: whether the sum ranges for 0 to K or from 1 to K is immaterial.

Multiple-layer neural networks. This idea can be generalized to multiple-layer neural networks. For example, the z_{ik}'s could be re-combined into new variable, say w_{im}, and so on...

This idea is at the core of deep learning methods.

Neural network applications. Neural nets have been used for time series forecasting in finance (see e.g. Kelly and Xiu, 2018), and are routinely used in the industry to perform pattern or speech recognition...

Computation

Gradient descent. The goal is to compute the solution to a problem of the form:

$$\underset{c}{\operatorname{argmin}}\, Q(c) = \sum_{i=1}^{N} (y_i - m(x_i, c))^2.$$

A simple idea is to use Newton's method. Starting at c_0 we compute:

$$c_1 = c_0 - [Q''(c_0)]^{-1} Q'(c_0).$$

Then we proceed iteratively using:

$$c_{s+1} = c_s - [Q''(c_s)]^{-1} Q'(c_s),$$

until c_s and c_{s+1} are close enough relative to a given distance measure (for example: the maximum of the absolute value difference across components is small enough).

This method is called Newton-Raphson. It requires not only knowledge of the gradient Q' but also of the derivative Q''.

There are alternative methods that do not rely on Q'', e.g. methods based on line search.

Computing derivatives: backpropagation. A key challenge in implementing gradient descent in neural networks is to compute Q' (one generally finds a way not to have to compute Q'').

A trick is to use the chain rule of derivation in a smart way. This is called "backpropagation". Implementing this efficiently may reduce the computational cost substantially.

Local optima. Neural network objectives are typically non-convex. Gradient descent methods will then often converge to local optima only, not the global optimum.

A strategy is often to specify the initial values in a way that the response function is approximately linear. However there is so far no theory for why starting from this region is desirable.

Read Farrell, Liang and Misra (2018).

Stochastic gradient descent

Computational cost in gradient descent? To assess the computational cost, it is useful to distinguish two challenging situations. In big data situations there are many observations (N is large). In big models situations there are many parameters (p is large).

When both N and p are large, as in neural network applications, it is useful to find ways to work with a smaller sample.

Batch methods. In a batch method we use an update rule that only depends on a subsample of the data. For example, we can implement Newton's step using a subsample to compute Q.

Robbins and Monro (1951) proposed to use different, randomly drawn subsamples in each iteration. This method is now called stochastic gradient descent.

The trade-off is an improved computational efficiency (much faster computations in big data situations), but the method is stochastic and depends on a key tuning parameter: the "learning rate".

Formally, we compute:

$$c_{s+1} = c_s - \mu_s Q'_s(c_s),$$

where Q_s is based on a subsample of the data.

The learning rate μ_s is such that it should tend to zero as s tends to infinity, but not too fast. If it tends too fast then we do not learn from all of our data. If it tends too slowly to zero then the method will converge very slowly and the computational advantages will be lost.

This trade-off has been studied theoretically, and default choices exist. However stochastic gradient descent methods are often sensitive to this choice.

Stochastic gradient descent is widely used to estimate ("train") neural networks.

Approximation property and issues in regularization and inference

Approximation. The first property we know about neural networks is that they have good approximation properties. Any suitable function (continuous) can be approximated uniformly well with a single-layer neural network with a sufficient number of nodes. This result is due to Hal White, the inventor of the White formula in econometrics.

Statistical properties: parametric model. Neural networks are nonlinear regression models. The asymptotic theory of nonlinear regression is well understood.

In a model of the form:

$$y_i = m(x_i, \theta) + u_i,$$

with $\mathbb{E}(u_i \,|\, x_i) = 0$, let:

$$\widehat{\theta} = \underset{c}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i - m(x_i, c))^2.$$

We have, under commonly used conditions:

$$\sqrt{N}(\widehat{\theta} - \theta) \overset{d}{\to} \mathcal{N}(0, V),$$

where V takes a "sandwich" form. In particular, its denominator involves the gradient of the function m with respect to θ.

Statistical properties: nonparametric model. At the same time, this theory abstracts from some of the key feature of neural networks.

-Neural networks are supposed to be "flexible" approximation. It is appealing to think of the true function $m(x)$ to be nonparametric.

-Neural networks have many parameters in most applications. This causes problems such as high variance, multi-collinearity, overfit, ...

-Given the nonlinear nature of the objective function it is often unrealistic to think we can compute its global minimum. Yet the standard theory assumes that one has computed the global minimum.

-There is also a problem of lack of identification or weak identification, since γ_{jk} is not identified when $a_k = 0$. See Andrews and Cheng (2012).

Developing a statistical theory that accounts for these features is still an open problem.

Chapter 8: Factor and matrix methods

1. **PCA and factor models**

2. **Matrix methods**

3. **Economic applications of factor methods**

PCA and factor models

Principal components analysis

Principal components. Consider a (possibly random) matrix X, e.g. a matrix of regressors. The dimensions of X are $N \times p$. From the SVD of X we have $X = UDV'$. Note that we can take U to be $N \times p$, D to be $p \times p$, and V to be $p \times p$. Let us order the diagonal elements of D (which are non-negative) as $d_1 \geq ... \geq d_p \geq 0$.

The columns of V are also the eigenvectors of $X'X = VD^2V'$. $v_1, ..., v_p$ are called the principal component directions of X.

The first eigenvector v_1 is such that Xv_1 has the largest variance among all (normalized) linear combinations of X. Note that $Xv_1 = d_1u_1$. u_1 is the first (normalized) principal component of X.

Likewise, u_2 is the second (normalized) principal component, and so on.

In many cases, the first few principal components explain most of the variation in X.

Principal components calculation. Suppose one wants to compute the first principal component of X. One possibility is to minimize:

$$\sum_{i=1}^{N}\sum_{j=1}^{p}(x_{ij} - \lambda_i f_j)^2,$$

subject to one normalization such as $(1/p)\sum_{j=1}^{p} f_j = 1$.

The solutions are $\lambda = d_1u_1$ (the first principal component, multiplied by the first singular value of X) and $f = v_1$ (the first principal component direction of X).

This expression can be generalized to other principal components.

Note that this least-squares representation may not be directly helpful for computation, since the objective function is not convex, and their exist efficient algorithms to perform SVD.

However, the equivalence between PCA and the least squares problem is very useful to study interesting extensions of PCA.

Principal components regression. In regression models, a common way to reduce the dimensionality of X is to perform a PCA on its columns, and extract the principal components $u_1, u_2, ...$

Then, one regresses y on $u_1, ..., u_K$, where $K \leq p$. This is PC regression.

When $K = p$, PC regression is simply OLS. When $K < p$ this is a regularized version of OLS.

As in Lasso and series, K is a tuning parameter that governs the amount of regularization/dimension reduction.

Since the u_i are orthogonal, regressing y on $u_1, ..., u_K$ is equivalent to K univariate regressions on u_1, u_2,..., and u_K.

Static and dynamic factor models

Static factor models. We are now going to apply PCA methods and PCA-related methods in a context where one has N units followed over T periods.

For example, we may have $N = 20$ portfolios with prices varying over $T = 500$ days.

Alternatively, we may observed $N = 1000$ individuals for $T = 10$ periods, as in many panel data applications.

The model is then:

$$y_{it} = \mu_i' f_t + u_{it},\ i = 1, ..., N,\ t = 1, ..., T,$$

where μ_i and f_t are $K \times 1$ (note: both are column vectors).

K is the number of factors.

For example, when $K = 1$ we have:

$$y_{it} = \mu_i f_t + u_{it},$$

where μ_i and f_t are scalar. This is a one-factor model.

We are going to treat $\mu_1, ..., \mu_N$ and $f_1, ..., f_T$ as parameters ("fixed-effects").

A challenge is that the number of parameters increases with the sample size.

Matrix form. Let K be the number of factors. In matrix form, let us construct the matrices F (which is $T \times K$) and M which is $N \times K$), by concatenating the f_t's and the

μ_i's. Let us also construct the matrices Y $(N \times T)$ and U $(N \times T)$, by concatenating the y_i's and the u_i's.

The factor model is then:

$$Y = MF' + U.$$

Normalization. At best we can consistently estimate MF'. If Ω is $K \times K$ orthogonal, then $MF' = M\Omega\Omega'F' = (M\Omega)(F\Omega)'$. Hence M and F are at best identified up to right-multiplication by an orthogonal matrix. Identification holds up to a choice of "rotation". A standard normalization is to impose $F'F = I_K$.

Static models: estimation.

$$\min_{\mu_1,...,\mu_N,f_1,...,f_T} \sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \mu_i' f_t)^2,$$

subject to a normalization, as indicated above.

The solution is PCA. In particular, the estimates of f_t are the principal component directions of Y. The estimates of μ_i are the principal components of Y, multiplied by the singular values of X.

Large factor models: properties. Bai and Ng (2002, Econometrica) study the properties of principal components estimates in a model where N and T tend to infinity simultaneously.

Intuitively, in this model it is necessary to have both large N (to learn about f_t) and large T (to learn about μ_i).

Using the normalization $F'F/T + I_k$, they find, for all factors $k = 1, ..., K$:

$$\sqrt{T}(\widehat{\mu}_{ik} - \mu_{ik}) \overset{d}{\to} \mathcal{N}(0, V_\mu),$$

and

$$\sqrt{N}(\widehat{f}_{tk} - f_{tk}) \overset{d}{\to} \mathcal{N}(0, V_f).$$

In these expressions, the true μ_i and f_t are suitably normalized. (recall the lack of identification in the absence of such normalization)

Moreover, the asymptotic convergence relies on some assumptions about the relative rates of growth of N and T. Bai and Ng's results hold when T and N are of the same order asymptotically, i.e. when T/N is constant in the limit.

Bai and Ng also propose a method to estimate the number of factors.

Weak dependence. The intuition behind Bai and Ng's theory is that $Y = MF' + U$, where MF' has rank K, and U are errors. Provided that u_{it} are sufficiently "weakly correlated" across both i and t, one can learn about M and F from Y.

Weak correlation over time could mean an MA process, or an AR process with autoregressive coefficient $|\rho| < 1$.

Weak correlation across units is harder to visualize. A possible dependence structure is obtained if one specifies a "spatial weights matrix" W (of dimensions $N \times 1$) such that w_{ij} represents the dependence of y_{it} on y_{jt}. We typically normalize the diagonal of W to have zero elements. A spatial AR(1) model is $y_t = Wy_t + \varepsilon_t$. Under suitable conditions of the singular values of W, such a process is weakly dependent across i.

Dynamic factor models. The static factor model makes no assumptions of μ_i or f_t.

This makes using for the model for prediction difficult: how to predict f_t in the future?

This limitation is particularly severe in macroeconomics, when using factor models to predict inflation, GDP, ...

In a dynamic factor model, we add a specification of the process $f_t \mid f_{t-1}, \ldots$. Typically, the process is specified as an AR(p) or Vector AR(p).

Estimation can be based on maximum likelihood. Estimators are typically consistent for fixed N as T tends to infinity under correct specifications.

Interactive fixed-effects model

Bai's model. Bai (2009) considers the model:

$$y_{it} = x_{it}'\beta + \mu_i' f_t + u_{it},$$

where μ_i, f_t are unrestricted.

This model has appeal when focusing on β, the "effect" of x on y, while attempting to control for a rich set of factors.

Bai's estimator is:

$$\min_{\beta,\mu_1,\ldots,\mu_N,f_1,\ldots,f_T} \sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - x_{it}'\beta - \mu_i' f_t)^2,$$

subject to a normalization as above.

The rationale for this is estimator the above equivalence between PCA and the least-squares formulation. Here we simply add a covariate.

This estimation problem is not convex. Bai's strategy is to iterate between ordinary least squares for β (given μ and f) and PCA for μ, f (given β).

This iteration will reach a local minimum of the objective function. In practice, it is recommended to use several starting values.

Finally, we need an estimate of the number K of factors. This can be based on the Bai and Ng information criteria.

Bai's model: properties. As in Bai and Ng (2002), Bai considers an asymptotic sequence where N, T tends to infinity simultaneously.

He shows that:

$$\sqrt{NT}(\widehat{\beta} - \beta) \overset{d}{\to} \mathcal{N}(0, V_\beta),$$

where the expression for V_β is available in the paper.

This is a nice result, since it shows that $\widehat{\beta}$ has a standard distribution.

Pesaran's model. There are other models related to Bai's interactive fixed-effects model in the literature. In particular, Pesaran (2006) considers a model of the form:

$$y_{it} = x_{it}'\beta + \mu_i' f_t + u_{it}, \quad x_{it} = H_i f_t + v_{it}.$$

The difference with Bai's model is that he assumes that the factors f_t also enter the x_{it}'s. This simplifies estimation (but requires making an additional assumption).

The estimator Pesaran proposes, which is called "common correlated effects", is very simple to implement.

Matrix methods

Read Chapter 7 in Hastie, Tibshirani and Wainwright's book.

Nuclear norm penalization

Nuclear norm of a matrix. Analog of ℓ^1 norm for vectors. $\|M\|_1 = \operatorname{Tr} D$, where $M = UDV'$ is the SVD of M. $\|M\|_1$ is the sum of singular values of M.

Matrix Lasso estimator.

$$\min_{m_{11},\ldots,m_{NT}} \|Y - M\|^2 + \lambda \|M\|_1.$$

Here, $\|Y - M\|^2 = \sum_{j=1}^J \sum_{k=1}^K (y_{jk} - m_{ij})^2$ is the squared Frobenius norm of $Y - M$.

An alternative expression is:

$$\|Y-M\|^2=\text{Tr}((Y-M)'(Y-M)).$$

The type of sparsity that is enforced here is that $\widehat{M}$ has low rank.

This is similar to PCA or factor analysis.

The matrix Lasso objective is convex. It turns out that the minimization of $\|Y-M\|^2+\lambda\|M\|_1$ with respect to M is a convex problem.

This result is not obvious. In fact, the nuclear norm is the dual of the spectral norm (which itself is a norm – it corresponds to the largest singular value of the matrix). It is convex since it is a norm.

This duality can be understood as a matrix analog of the duality between ℓ^1 and ℓ^∞ norms for vectors.

Computation. For computation, one can proceed in a similar way as for the cyclic coordinate descent iterative algorithm for Lasso.

However, in the case of:

$$\min_{m_{11},...,m_{NT}} \|Y-M\|^2+\lambda\|M\|_1,$$

we have a nice explicit result due to Cai, Candes and Shen (2008):

$$\widehat{M}=UD_\lambda V',$$

where $Y=UDV'$ is the SVD of Y, and the elements of the diagonal D_λ are $d_j(\lambda)=(d_j-\lambda)^+$.

$\widehat{M}$ is obtained by application of the "singular value shrinkage operator".

Application to interactive fixed-effects models. Bai (2009) considers the model:

$$y_{it}=x_{it}'\beta+\mu_i'f_t+u_{it},$$

where μ_i, f_t are unrestricted.

This model has appeal when focusing on β, the "effect" of x on y, while attempting to control for a rich set of factors.

Bai's estimator is:

$$\min_{\beta,\mu_1,\ldots,\mu_N,f_1,\ldots,f_T} \sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - x_{it}'\beta - \mu_i' f_t)^2,$$

subject to a normalization as above.

This estimation problem is not convex. Bai's strategy is to iterate between ordinary least squares for β (given μ and f) and PCA for μ, f (given β).

A convex relaxation is obtained by using nuclear norm regularization, as studied in Moon and Weidner (2019).

Matrix completion

Matrix completion using nuclear norm. Suppose you know some entries of a matrix M, say $m_{ij} = z_{ij}$ for (i, j) in a set Ω.

Matrix completion aims at solving the problem of finding a matrix with minimum rank that satisfies $m_{ij} = z_{ij}$ for (i, j) in Ω.

Since this problem is non-convex, we use the nuclear norm to "convexify" it.

The problem is then:

$$\min_{M} \|M\|_1, \text{ s.t. } m_{ij} = z_{ij} \text{ for all } (i, j) \in \Omega.$$

Example 1: the Netflix movie challenge. Read Chapter 7 in Hastie, Tibshirani and Wainwright.

Netflix launched a competition in 2006. The netflix dataset has $N = 17770$ movies and $T = 480189$ customers. The data contains ratings of movies. However the missing data problem is very severe: less than 1% of the ratings are present.

The goal is to predict the ratings for the unrated movies.

A possible objective is to minimize the rank of M (which is $N \times T$), subject to $m_{ij} = z_{ij}$ for the available entries $(i, j) \in \Omega$.

Such a "low-rank heuristic" provides good prediction results for the netflix challenge.

Example 2: panel data models. Read Athey, Bayati, Doutchenko, Imbens and Khosravi (2018).

Economic applications of factor methods

Macroeconomic forecasting

Stock and Watson's dynamic factor models. Sargent and Sims (1977) showed that two dynamic factors could explain a large fraction of the variance of important U.S. quarterly macroeconomic variables, including output, employment, and prices.

In a series of papers, Stock and Watson extended these approaches to jointly model a large number of macroeconomic time series, and use dynamic factor models for forecasting purposes. Stock and Watson (2011, Oxford handbook on economic forecasting) provides a survey of this line of work.

Stock and Watson (2012, NBER): a new factor in the Great recession. Stock and Watson (2012) analyze the macroeconomic dynamics of the 2007-09 recession in the US and the subsequent slow recovery. They use a dynamic factor model with 200 variables. They find that the Great recession was associated with the emergence of a new – financial – factor.

Generalizing differences-in-differences

Traditional Diff in Diff. Panel data differences-in-differences is a leading framework for empirical work in economics. The model is:

$$y_{it} = x_{it}'\beta + \mu_i + f_t + u_{it}.$$

The parameters μ_i, f_t are unrestricted. However, unlike in the interactive fixed-effects model, these parameters enter the model additively.

The key assumption in diif-in-diff is that, when $x_{it} = x$, $x'\beta + \mu_i + f_t$ is a common, parallel time pattern across individuals. This is referred to as the "common pre-trends assumption".

The standard estimator relies on double differencing: take out the mean over t and the mean over i, and add the grand mean.

Diff in Diff with factors. Using the methods in this chapter, we are able to relax the common pre-trends assumption.

Indeed consider Bai's model:

$$y_{it} = x_{it}'\beta + \mu_i' f_t + u_{it},$$

or Bai's model with additional additive parameters (with suitable normalizations):

$$y_{it} = x_{it}'\beta + \mu_i' f_t + a_i + b_t + u_{it}.$$

These models are easy to estimate using Bai's method. This can help relax the key assumption in diff-in-diff.

Another generalization of diff in diff: synthetic control. Read Adadie, Diamond and Hainmueller (2010).

Abadie's synthetic control method is closely related to factor models.

Indeed, in their JASA paper they use a one-factor model (with $K = 1$) to motivate their synthetic control estimator.

An alternative to synthetic control is to estimate a one-factor (or multiple-factor) model, as we have learned in this chapter.

Chapter 9: Latent variables

1. **Fixed-effects and random-effects**
2. **Grouped fixed-effects using kmeans clustering**
3. **Variational inference, with applications**

Fixed-effects and random-effects

Hierarchical models and likelihoods

Static and dynamic panel data.

Panel data techniques have many applications: in labor economics (workers, firms...), trade (firms, countries, industries, products...), industrial organization (markets, products...), or cross-country macroeconomics (countries, sectors...).

Two specific features of panel data are (1) the presence of unobserved heterogeneity, and (2) the dynamic relationships involved.

We face two main challenges.

- We have often little knowledge about the unobserved heterogeneity. For example, economic models are typically silent about it.
- These model are often high-dimensional, due to the fact that we need to model the distribution of variables across individuals and over time.

Likelihood functions

Conditional likelihood. Let $Y_i = (Y_{i1}, ..., Y_{iT})'$ denotes a sequence of outcomes.

Let $X_i = (X_{i1}', ..., X_{iT}')'$ denote a sequence of covariates, which for the moment we assume are strictly exogenous.

α_i denotes a vector of unobserved individual effects (it does not need to be scalar).

We will focus on likelihood models, where the distribution of Y_i given X_i and α_i is parametric, indexed by θ:

$$f\left(Y_i | X_i, \alpha_i; \theta\right).$$

The (conditional) log-likelihood function is thus

$$L(\theta) = \sum_{i=1}^{N} \ln\left(f\left(Y_i|X_i, \alpha_i; \theta\right)\right).$$

It is important to note that, in this conditional approach, no assumptions are imposed on the individual fixed-effects α_i. We will see in the next Chapter that this differs from the random-effects approach, which specifies the cross-sectional distribution of α_i.

Dynamic models. Dynamic models are covered by this formulation, as

$$f(Y_i|X_i, \alpha_i; \theta) = \prod_{t=1}^{T} f(Y_{it}|Y_i^{t-1}, X_i, \alpha_i; \theta),$$

where $Y_i^t = (Y_{it}, Y_{i,t-1}, ...)$, and X_i contains strictly exogenous regressors and initial conditions.

Dealing with predetermined covariates can be done be writing the full likelihood function of (Y_i, X_i) conditional on α_i and initial conditions (Y_{i0}, X_{i0}) (which here we assume to be observed: this is a simple matter of notation). This requires that the researcher be willing to parametrically specify the conditional distribution of X_{it} given X_i^{t-1}, Y_i^{t-1}, and α_i (the "feedback process").

Example 1. Consider the dynamic Gaussian autoregressive model:

$$Y_{it} = \rho Y_{i,t-1} + X_{it}'\beta + \alpha_i + \varepsilon_{it}, \tag{1}$$

where $\varepsilon_{it} \sim \mathcal{N}(0, \sigma^2)$, i.i.d. across individuals and time.

Letting $\theta = (\rho, \beta, \sigma^2)'$, the conditional likelihood function is

$$f(Y_i|X_i, Y_{i0}, \alpha_i; \theta) = \frac{1}{\sigma^T} \prod_{t=1}^{T} \phi\left(\frac{Y_{it} - \rho Y_{i,t-1} - X_{it}'\beta - \alpha_i}{\sigma}\right),$$

where ϕ is the standard normal pdf.

Example 2 Consider the dynamic probit model:

$$Y_{it} = \mathbf{1}\left\{\rho Y_{i,t-1} + X_{it}'\beta + \alpha_i + \varepsilon_{it} \geq 0\right\},$$

where $\varepsilon_{it} \sim \mathcal{N}(0, 1)$, i.i.d. across individuals and time.

In this case the conditional likelihood function, characterized by $\theta = (\rho, \beta)'$, is

$$\begin{aligned} f(Y_i|X_i, Y_{i0}, \alpha_i; \theta) &= \prod_{t=1}^{T} \Phi\left(\rho Y_{i,t-1} + X_{it}'\beta + \alpha_i\right)^{Y_{it}} \\ &\quad \times \left[1 - \Phi\left(\rho Y_{i,t-1} + X_{it}'\beta + \alpha_i\right)\right]^{1-Y_{it}}, \end{aligned}$$

where Φ is the standard normal cdf.

Three approaches

Read Arellano and Bonhomme (2009).

Fixed-effects. The fixed-effects estimator maximizes the conditional log-likelihood function with respect to all parameters, including the individual-specific α_i.

Given an i.i.d. sample $(Y_i', X_i')'$, $i = 1, ..., N$.

$$(\widehat{\theta}^{FE}, \widehat{\alpha}_1^{FE}, ..., \widehat{\alpha}_N^{FE}) = \underset{(\theta, \alpha_1, ..., \alpha_N)}{\operatorname{argmax}} \sum_{i=1}^{N} \ln f(Y_i|X_i, \alpha_i; \theta).$$

Concentrating out α_i yields:

$$\widehat{\theta}^{FE} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \ln f(Y_i|X_i, \widehat{\alpha}_i^{FE}(\theta); \theta),$$

where:

$$\widehat{\alpha}_i^{FE}(\theta) = \underset{\alpha_i}{\operatorname{argmax}} \ln f(Y_i|X_i, \alpha_i; \theta).$$

Note that $\widehat{\alpha}_i^{FE}(\theta)$ is estimated based on T observations.

Random-effects. Let $Y_i = (Y_{i1}, ..., Y_{iT})'$ denote the full sequence of outcomes, and let $X_i = (X_{i1}', ..., X_{iT}')'$ denote a sequence of strictly exogenous covariates.

The likelihood of Y_i is conditioned on X_i and the vector of individual effects α_i is assumed to belong to a parametric family $f\left(Y_i|X_i, \alpha_i; \theta\right)$.

In a random-effects fashion, the researcher will complete the model by specifying a parametric distribution for the individual effects, conditional on exogenous covariates and initial conditions. Let $f\left(\alpha_i|X_i; \xi\right)$ denote that distribution, which is fully characterized by the parameter ξ.

A popular example is to specify α_i to be Gaussian with a mean that is a linear combination of exogenous covariates, and a constant variance, yielding:

$$f_{\alpha|x}\left(\alpha_i|X_i;\xi\right)=\frac{1}{\nu}\phi\left(\frac{\alpha_i-X_i'\mu}{\nu}\right),$$

where $\xi=(\mu,\nu)$.

Chamberlain (1984) introduces this specification in the static probit model. Alvarez and Arellano (2003) use a similar specification for the distribution of individual effects of an autoregressive model where the conditional mean of α_i is linear in the initial condition of the process.

Once a distribution has been postulated for the individual effects, the researcher will base inference on the average (or integrated) likelihood:

$$f(Y_i|X_i;\theta,\xi)=\int f(Y_i|X_i,\alpha;\theta)f\left(\alpha|X_i;\xi\right)d\alpha, \tag{2}$$

where the integral is taken over the support of the distribution of individual effects (typically the real line when α_i is scalar).

Note that the integrated likelihood function is fully characterized by the parameter (θ,ξ) so that, under correct specification, a parametric approach can be used for estimation and inference.

Bayesian. The average likelihood function given by (2) is also appealing from a Bayesian perspective. A Bayesian researcher would start by specifying a joint prior distribution for $(\alpha_1,...,\alpha_N,\theta)$.

Viewing $\alpha_1,...,\alpha_N$ as an i.i.d. sample of missing data, it is natural to assume prior conditional independence of $\alpha_1,...,\alpha_N$ given θ. Under this assumption, the joint prior conditioned on covariates can be decomposed as:

$$\pi\left(\alpha_1,...,\alpha_N,\theta\right)=\pi_1\left(\alpha_1|\theta\right)\times...\times\pi_N\left(\alpha_N|\theta\right)\times\pi\left(\theta\right).$$

In this case the posterior distribution for θ is proportional to:

$$\begin{aligned}p\left(\theta|Y_1,...,Y_N,X_1,...,X_N\right) \quad &\propto \quad \pi\left(\theta\right)\int f(Y_1|X_1,\alpha_1;\theta)\pi_1\left(\alpha_1|\theta\right)d\alpha_1\times...\\ &\qquad ...\times\int f(Y_N|X_N,\alpha_N;\theta)\pi_N\left(\alpha_N|\theta\right)d\alpha_N.\end{aligned} \tag{3}$$

Therefore, the random-effects integrated likelihood (2) can be interpreted in a Bayesian perspective as a marginal likelihood, where the (hierarchical) prior specification on individual effects is given by:

$$\pi_i\left(\alpha_i|\theta;\xi\right)=f\left(\alpha_i|X_i;\xi\right).$$

Random-effects specifications are a special case of hierarchical Bayesian approaches, where the prior distribution of individual effects is assumed independent of common parameters (but not of covariates).

Intuitions from large-N,T asymptotics. As T increases random-effects estimators become consistent, irrespective of the form of the postulated distribution of individual effects (Arellano and Bonhomme, 2009). There properties are thus qualitatively similar to those of fixed-effects estimators.

The reason for consistency is that (taking a dynamic model for concreteness)

$$\log f(Y_i|X_i,\alpha;\theta)=\sum_{t=1}^{T}\log f(Y_{it}|Y_i^{t-1},X_i,\alpha;\theta)$$

is a sum of T time-series observations, so the effect of the prior distribution $f(\alpha_i|X_i;\xi)$ becomes negligible compared to that of the likelihood as the number of time periods increases.

More formally,

$$\begin{aligned}\int f(Y_i|X_i,\alpha;\theta)f\left(\alpha|X_i;\xi\right)d\alpha &= \int e^{\sum_{t=1}^{T}\ln f(Y_{it}|X_{it},\alpha;\theta)+\ln f(\alpha|X_i;\xi)}d\alpha\\ &\approx \int e^{\sum_{t=1}^{T}\ln f(y_{it}|x_{it},\alpha;\theta)}d\alpha\\ &\approx f(Y_i|X_i,\widehat{\alpha}_i^{FE}(\theta);\theta).\end{aligned}$$

For small T, however, the estimator suffers from a bias of order $1/T$ under standard regularity conditions. This is also similar to fixed-effects.

Both the fixed-T inconsistency of fixed-effects and random-effects maximum likelihood approaches may thus be viewed as a manifestation of the same incidental parameter problem.

Grouped fixed-effects using kmeans clustering

Read Bonhomme, Lamadon and Manresa (2017).

K-means for dimension reduction

Consider a panel data setup, where we denote outcome variables and exogenous covariates as $Y_i = (Y_{i1}', ..., Y_{iT}')'$ and $X_i = (X_{i1}', ..., X_{iT}')'$, respectively, for $i = 1, ..., N$.The conditional density of Y_i given X_i is given by:

$$\ln f_i(\alpha_{i0}, \theta_0) = \sum_{t=1}^{T} \ln f(Y_{it} \,|\, Y_{i,t-1}, X_{it}, \alpha_{i0}, \theta_0), \tag{4}$$

and the density of exogenous covariates X_i take the form:

$$\ln g_i(\mu_{i0}) = \sum_{t=1}^{T} \ln g(X_{it} \,|\, X_{i,t-1}, \mu_{i0}).$$

We leave the form of g unrestricted, and in estimation we will use a conditional likelihood approach based on f_i alone. In other words, in applications the researcher only needs to specify the parametric form of $f_i(\alpha_{i0}, \theta_0)$ in (4). However, the dimension of μ_{i0} will play an important role when studying the properties of two-step GFE.

Rich patterns of heterogeneity. Unobserved heterogeneity can vary over time. Variation in unobservables over time (e.g., over the business cycle), age (over the life cycle), or markets (over counties or MSA) is of interest in many applications. In this case conditional densities take the form:

$$\ln f_i(\alpha_{i0}, \theta_0) = \sum_{t=1}^{T} \ln f(Y_{it} \,|\, Y_{i,t-1}, X_{it}, \alpha_{it0}, \theta_0); \quad \ln g_i(\mu_{i0}) = \sum_{t=1}^{T} \ln g(X_{it} \,|\, X_{i,t-1}, \mu_{it0}),$$

where $\alpha_{i0} = (\alpha_{i10}', ..., \alpha_{iT0}')'$ and $\mu_{i0} = (\mu_{i10}', ..., \mu_{iT0}')'$.

Two-step estimation. We rely on the individual-specific moments h_i to learn about the individual types ξ_{i0}. Specifically, we partition the individual units into K groups, corresponding to group indicators $\widehat{k}_i \in \{1, ..., K\}$ that approximate the moments h_i in the following sense:

$$\left(\widehat{h}, \widehat{k}_1, ..., \widehat{k}_N\right) = \underset{\left(\widetilde{h}, k_1, ..., k_N\right)}{\operatorname{argmin}} \sum_{i=1}^{N} \left\| h_i - \widetilde{h}(k_i) \right\|^2, \tag{5}$$

where $\{k_i\}$ are partitions of $\{1, ..., N\}$ into at most K groups, and $\widetilde{h} = (\widetilde{h}(1)', ..., \widetilde{h}(K)')'$, where $\widetilde{h}(k)$ are vectors. Note that $\widehat{h}(k)$ is simply the mean of h_i in group $\widehat{k}_i = k$.

The optimization problem in (5) is referred to as *kmeans* in machine learning and computer science. In (5) the minimum is taken with respect to all possible partitions $\{k_i\}$.

Computing a global minimum may be challenging, yet fast and stable heuristic algorithms such as Lloyd's algorithm have been developed; see Bonhomme and Manresa (2015) for references. In the asymptotic analysis, following the literature since Pollard (1981, 1982), we will focus on the properties of the global minimum in (5) and abstract from optimization error. Lastly, note that the quadratic loss function in (5) can accommodate weights on different components of h_i, although for simplicity we present the unweighted case.

In the second step we maximize the log-likelihood function with respect to common parameters and group-specific effects, where the groups are given by the $\widehat{k}_i$ estimated in the first step. We define the two-step GFE estimator as:

$$\left(\widehat{\theta}, \widehat{\alpha}\right) = \underset{(\theta,\alpha)}{\operatorname{argmax}} \sum_{i=1}^{N} \ln f_i\left(\alpha\left(\widehat{k}_i\right), \theta\right), \tag{6}$$

where the maximization is with respect to θ and $\alpha = (\alpha(1)', ..., \alpha(K)')'$, with $\alpha(k)$ being parameter vectors. Note that, in contrast with fixed-effects maximum likelihood, this second step involves a maximization with respect to K group-specific parameters instead of N individual-specific ones. In models with time-varying heterogeneity, $\alpha(k)$ will simply be a vector $(\alpha_1(k)', ..., \alpha_T(k)')'$.

Properties of K-means

Discrete DGP. Read Bonhomme and Manresa (2015).

$$y_{it} = x_{it}'\beta_0 + \alpha(k_{i0}, t) + u_{it}.$$

The population consists of K groups. BM provide conditions for perfect group classification. The algorithm is a simple modification of Lloyd's algorithm for kmeans.

Application: income and paths of democracy based on Acemoglu *et al.* (2008).

Inference post model selection? The results in BM assume that groups can be approximated uniformly well in large samples. Under these conditions the asymptotic distribution of the second-step estimator coincide with those of the estimator where the population groups are known ("oracle case"). However this falls under the Leeb and Potscher critique: this asymptotic scenario may not mimic finite-sample situations well.

Approximating continuous heterogeneity.

Assumption 1 *(underlying dimension)*
(a) Time-invariant heterogeneity: There exist vectors ξ_{i0} of dimension d, and two Lipschitz-continuous functions α and μ, such that $\alpha_{i0} = \alpha(\xi_{i0})$ and $\mu_{i0} = \mu(\xi_{i0})$.

(b) Time-varying heterogeneity: There exist vectors ξ_{i0} of dimension d, vectors λ_{t0} of dimension d_λ, and two functions α and μ that are Lipschitz-continuous in their first arguments, such that $\alpha_{it0} = \alpha(\xi_{i0}, \lambda_{t0})$ and $\mu_{it0} = \mu(\xi_{i0}, \lambda_{t0})$.

Assumption 2 *(injective moments)*
There exist vectors h_i of fixed dimension, and a Lipschitz-continuous function φ, such that $\text{plim}_{T\to\infty} h_i = \varphi(\xi_{i0})$, and $\frac{1}{N}\sum_{i=1}^N \|h_i - \varphi(\xi_{i0})\|^2 = O_p(1/T)$ as N, T tend to infinity. Moreover, there exists a Lipschitz-continuous function ψ such that $\xi_{i0} = \psi(\varphi(\xi_{i0}))$.

Rates, as a function of underlying dimension.

Lemma 3 *Let Assumption 2 hold. Then, as N, T, K tend to infinity we have:*

$$\frac{1}{N}\sum_{i=1}^N \left\| \widehat{h}(\widehat{k}_i) - \varphi(\xi_{i0}) \right\|^2 = O_p\left(\frac{1}{T}\right) + O_p\left(B_\xi(K)\right).$$

Lemma 4 *(Graf and Luschgy, 2002) Suppose that ξ_{i0} are random vectors with a distribution whose support is compact in $\mathbb{R}^d$. Then, as N, K tend to infinity we have $B_\xi(K) = O_p(K^{-\frac{2}{d}})$.*

For example, Lemma 4 implies that $B_\xi(K) = O_p(K^{-2})$ when ξ_{i0} is one-dimensional, and $B_\xi(K) = O_p(K^{-1})$ when ξ_{i0} is two-dimensional.

Theorem under time-invariant heterogeneity:

Theorem 5 *As N, T, K tend to infinity we have:*

$$\widehat{\theta} = \theta_0 + H^{-1}\frac{1}{N}\sum_{i=1}^N s_i + O_p\left(\frac{1}{T}\right) + O_p\left(K^{-\frac{2}{d}}\right) + o_p\left(\frac{1}{\sqrt{NT}}\right). \tag{7}$$

Theorem under time-varying heterogeneity:

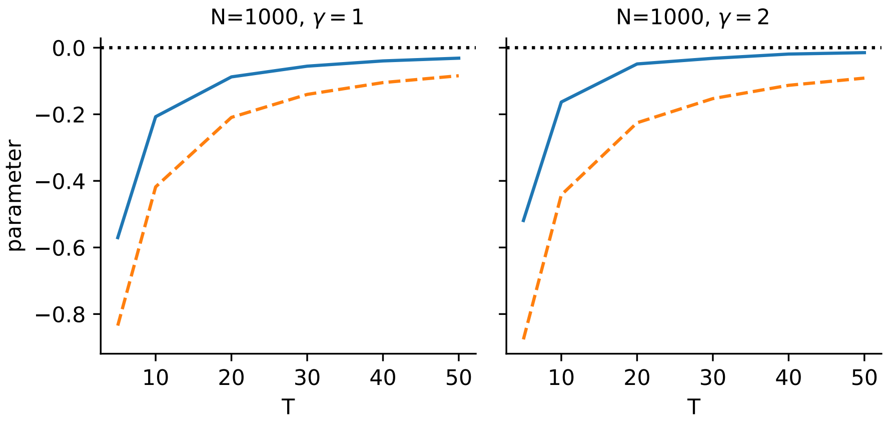
Theorem 6 *As N, T, K tend to infinity such that K/N tends to zero, we have:*

$$\widehat{\theta} = \theta_0 + H^{-1}\frac{1}{N}\sum_{i=1}^N s_i + O_p\left(\frac{1}{T}\right) + O_p\left(\frac{K}{N}\right) + O_p\left(K^{-\frac{2}{d}}\right) + o_p\left(\frac{1}{\sqrt{NT}}\right). \tag{8}$$

An example: model of wages and labor force participation. Consider first the following model of wages W_{it}^* and labor force participation Y_{it}:

$$\left\{ \begin{array}{rcl} Y_{it} & = & \mathbf{1}\left\{u(\alpha_{i0}) \geq c(Y_{i,t-1}; \theta_0) + U_{it}\right\}, \\ W_{it}^* & = & \alpha_{i0} + V_{it}; \quad W_{it} \, = \, Y_{it}W_{it}^*, \end{array} \right. \tag{9}$$

Figure 1: Model of wages and participation



Notes: GFE in solid, FE in dashed. Averages among 1000 simulations. N=1000 and T is indicated in the graph. γ is the risk aversion.

where the wage is only observed when i works, U_{it} are i.i.d. standard normal, independent of the past Y_{it}'s and α_{i0}, and V_{it} are i.i.d. independent of all U_{it}'s, Y_{i0} and α_{i0}. Here the same payoff α_{i0}, unobserved to the econometrician, drives the wage and the decision to work. Individuals have common preferences over payoffs denoted by the utility function u, and the cost function is state dependent.

We focus on the difference in costs $c(0;\theta_0) - c(1;\theta_0)$, which reflects the presence of state dependence in participation decisions. We take the CRRA utility specification:

$$u(\alpha) = \frac{e^{\alpha(1-\gamma)} - 1}{1-\gamma},$$

and set $c(0;\theta_0) = 0$ and $c(1;\theta_0) = -1$. We consider two values for risk aversion: $\gamma \in \{1, 2\}$. We model the initial condition as $Y_{i0} = \mathbf{1}\{u(\alpha_i) \geq c(1;\theta_0) + U_{i0}\}$, with U_{i0} standard normal. As moments for GFE we take $\overline{W}_i$ and $\overline{Y}_i$.

Variational inference, with applications

The variational approach

Setup. Consider a likelihood $f(Y_i \,|\, X_i, \alpha_i, \theta)$, $f_{i\theta}(\alpha_i)$ for short.

-$Y_i = (Y_{i1}, ..., Y_{iT})$ are outcomes, $X_i = (X_{i1}, ..., X_{iT})$ are exogenous covariates (and initial conditions in dynamic models).

-α_i is unobserved heterogeneity.

Consider a correlated random-effects specification $g(\alpha_i \,|\, X_i, \gamma)$ for α_i, $g_{i\gamma}(\alpha_i)$ for short.

The **random-effects estimator** maximizes the log-integrated likelihood:

$$L^{\text{RE}}(\theta, \gamma) = \sum_{i=1}^{n} \ln \int f_{i\theta}(\alpha_i) g_{i\gamma}(\alpha_i) d\alpha_i.$$

Variational approximation. Denote the **posterior distribution** of α_i as:

$$p_{i\theta,\gamma}(\alpha_i) = \frac{f_{i\theta}(\alpha_i) g_{i\gamma}(\alpha_i)}{\int f_{i\theta}(\alpha) g_{i\gamma}(\alpha) d\alpha}.$$

In the variational approach, we approximate $p_{i\theta,\gamma}(\alpha_i)$ by a **parametric distribution** $q_{\eta_i}(\alpha_i)$ with parameter η_i.

The **variational random-effects (VRE)** estimator maximizes the lower bound:

$$L^{\text{VRE}}(\theta, \gamma, \eta_1, ..., \eta_N) = \sum_{i=1}^{n} \ln \int f_{i\theta}(\alpha_i) g_{i\gamma}(\alpha_i) d\alpha_i - \mathbb{E}_{q_{\eta_i}} \ln \frac{q_{\eta_i}(\alpha_i)}{p_{i\theta,\gamma}(\alpha_i)}.$$

-Since $\mathbb{E}_{q_{\eta_i}} \ln \frac{q_{\eta_i}(\alpha_i)}{p_{i\theta,\gamma}(\alpha_i)}$ is the KL divergence between the approximating and true posteriors, VRE and RE coincide when the family $q_{\eta_i}(\alpha_i)$ is "very rich".

A **key property** is that (for $\eta = (\eta_1, ..., \eta_N)$):

$$\begin{aligned} L^{\text{VRE}}(\theta, \gamma, \eta) &= \sum_{i=1}^{n} \ln \int f_{i\theta}(\alpha_i) g_{i\gamma}(\alpha_i) d\alpha_i - \mathbb{E}_{q_{\eta_i}} \ln \frac{q_{\eta_i}(\alpha_i)}{p_{i\theta,\gamma}(\alpha_i)} \\ &= \sum_{i=1}^{n} \mathbb{E}_{q_{\eta_i}} \ln f_{i\theta}(\alpha_i) g_{i\gamma}(\alpha_i) - \mathbb{E}_{q_{\eta_i}} \ln q_{\eta_i}(\alpha_i). \end{aligned}$$

The last expression no longer involves the log of the integrated likelihood.

Given a well-chosen family $q_{\eta_i}(\alpha_i)$, the expectations of log-densities can be computed efficiently.

However, an obvious drawback of the approach is that we are no longer maximizing the "right" objective function.

VRE with a point-mass approximation. Suppose that $q_{\eta_i}(\alpha_i)$ is the **point mass density** at η_i.

In this case:

$$\begin{aligned} L^{\text{VRE}}(\theta,\gamma,\eta) &= \sum_{i=1}^{n} \mathbb{E}_{q_{\eta_i}} \ln f_{i\theta}(\alpha_i) g_{i\gamma}(\alpha_i) - \mathbb{E}_{q_{\eta_i}} \ln q_{\eta_i}(\alpha_i) \\ &= \sum_{i=1}^{n} \ln f_{i\theta}(\eta_i) g_{i\gamma}(\eta_i). \end{aligned}$$

The VRE of θ is the **fixed-effects** maximum a posteriori estimator.

It is first-order biased as n and T tend to infinity, even when the random-effects distribution $g_{i\gamma}(\alpha_i)$ is correctly specified.

Gaussian VRE. Suppose now that $\eta_i = (\mu_i, \Sigma_i)$, and $q_{\mu_i,\Sigma_i}(\alpha_i)$ is the **normal density** with mean μ_i and variance Σ_i.

Using the "reparameterization trick", we have:

$$\mathbb{E}_{q_{\mu_i,\Sigma_i}} \ln f_{i\theta}(\alpha_i) g_{i\gamma}(\alpha_i) = \mathbb{E}_{\varepsilon_i} \ln[f_{i\theta} g_{i\gamma}] \left(\mu_i + \Sigma_i^{\frac{1}{2}} \varepsilon_i\right),$$

where $\varepsilon_i \sim \mathcal{N}(0, I)$. This will allow us to approximate the expectation by simulation.

Moreover:

$$\mathbb{E}_{q_{\mu_i,\Sigma_i}} \ln q_{\mu_i,\Sigma_i}(\alpha_i) = -\frac{1}{2} \ln \det \Sigma_i + C,$$

where C is a constant.

Hence, up to irrelevant constants:

$$L^{\text{VRE}}(\theta,\gamma,\mu,\Sigma) = \sum_{i=1}^{n} \mathbb{E}_{\varepsilon_i} \ln[f_{i\theta} g_{i\gamma}] \left(\mu_i + \Sigma_i^{\frac{1}{2}} \varepsilon_i\right) + \frac{1}{2} \ln \det \Sigma_i.$$

A **simulated counterpart** using nJ standard normal ε draws is then:

$$\widetilde{L}^{\text{VRE}}(\theta,\gamma,\mu,\Sigma) = \sum_{i=1}^{n} \frac{1}{J} \sum_{j=1}^{J} \ln[f_{i\theta} g_{i\gamma}] \left(\mu_i + \Sigma_i^{\frac{1}{2}} \varepsilon_{ij}\right) + \frac{1}{2} \ln \det \Sigma_i.$$

Hence the VRE objective is a **modified fixed-effects** log-likelihood. Computation can take advantage of automatic differentiation and stochastic gradient descent.

Random-effects: asymptotics. When n and T tend to infinity at the same rate, the **random-effects** estimator $\widehat{\theta}^{\text{RE}}$ behaves as follows (Arellano and Bonhomme, 2009):

-Under correct specification of $g_{i\gamma}(\alpha_i)$:

$$\sqrt{nT}\left(\widehat{\theta}^{\text{RE}} - \theta_0\right) \xrightarrow{d} \mathcal{N}(0, V_\theta).$$

-Under incorrect specification of $g_{i\gamma}(\alpha_i)$:

$$\sqrt{nT}\left(\widehat{\theta}^{\text{RE}} - \theta_0 - B/T\right) \xrightarrow{d} \mathcal{N}(0, V_\theta),$$

where B is a constant.

Posterior distribution: asymptotics. Consider now the **posterior distribution** $p_{i\theta,\gamma}(\alpha_i)$.

As T tends to infinity and under sufficient regularity, the Bernstein von Mises theorem implies that $p_{i\theta,\gamma}(\alpha_i)$ is approximately:

$$\mathcal{N}(\widehat{\alpha}_i(\theta), \Omega_i(\theta)/T),$$

where:

-$\widehat{\alpha}_i(\theta)$ is the fixed-effects estimator of α_i for fixed θ.

-The variance $\Omega_i(\theta)/T$ may not coincide with the asymptotic variance of $\widehat{\alpha}_i(\theta)$ (Kleijn and van der Vaart, 2012).

Variational random-effects: asymptotic equivalence. We have, for large n, T:

$$\begin{aligned}\frac{1}{n}L^{\text{RE}}(\theta,\gamma) - \max_{\mu,\Sigma}\frac{1}{n}L^{\text{VRE}}(\theta,\gamma,\mu,\Sigma) &= \min_{\mu,\Sigma}\ \frac{1}{n}\sum_{i=1}^{N}\mathbb{E}_{q_{\mu_i,\Sigma_i}}\ln\frac{q_{\mu_i,\Sigma_i}(\alpha_i)}{p_{i\theta,\gamma}(\alpha_i)} \\ &= o_p(T^{-1}),\end{aligned}$$

since $p_{i\theta,\gamma}(\alpha_i)$ can be approximated to first order by the Gaussian distribution $\mathcal{N}(\widehat{\alpha}_i(\theta), \Omega_i(\theta)/T)$.

Hence, when n and T tend to infinity at the same rate we will have the **asymptotic equivalence**:

$$\sqrt{nT}\left(\widehat{\theta}^{\text{VRE}} - \widehat{\theta}^{\text{RE}}\right) = o_p(1).$$

A similar result will hold for the simulated VRE estimator, for any fixed $J > \dim \mu_i + \dim \Sigma_i$.

Network models of link formation

Network formation. Active literature. Friendship, links between firms, banks... Logit models (Graham). Models with network externalities (Mele).

Stochastic blockmodel.

$$D_{ij} = \mathbf{1}\{U(X_{ij}, A_i, A_j, \varepsilon_{ij}) > 0\},$$

where ε_{ij} are i.i.d. across (i, j) pairs. More general than fixed-effects models.

Maximum likelihood? Intractable likelihood objective.

Variational approach. Postulate a model that is independent across pairs. Mean field approximation. Optimization. There are statistical guarantees in this context (Bickel and co-authors, 2013).

Random effects for metwork formation. Consider the **network formation** model in Graham (2017):

$$Y_{ij} = \mathbf{1}\left\{X_{ij}'\theta + \alpha_i + \alpha_j + \varepsilon_{ij} \geq 0\right\},$$

where ε_{ij} are i.i.d. standard logistic.

Model $\alpha_i \sim iid\mathcal{N}(\overline{X}_i'\gamma_1, \gamma_2)$. The RE estimator is **not tractable**.

The VRE estimator is again similar as before, and we again expect first-order asymptotic equivalence with RE.

VRE can also be used:

-To estimate average effects (such as an average marginal effect of X_{ij} on link formation).

-To relax the logistic specification, e.g. using a neural network model.

Text analysis

Text data. Increasingly popular. Read Gentzkow and Shapiro, “Text as Data”.

Latent Dirichlet allocation. A document is a mixture of topics. Goal: recover topic distributions. Example: Blei and Lafferty’s analysis of *Science* magazine.

The LDA model. Hierarchical model.

Maximum likelihood? Intractable likelihood objective.

Variational approach. Mean field approximation. Optimization. In this case statistical guarantees are not obvious.

An alternative: non-negative matrix factorization (NMF).