Highlights of Selected Papers (Comprehensive '25)

Tanishq Kumar Prasad

March 2025

This document only contains the main results and ideas from the selected papers. For looking at the proofs of various results, and the simulation studies/data analytic experiments done, please use the drive links to view the paper/slides.

Prologue

A short note on LARS (needed for Supratim and Swagato's papers)

Least Angle Regression is an algorithm to estimate the regression parameters, ultimately yielding the least squares estimate. However, the key objective is to provide a "solution path" - ie to show how the regression coefficients evolve with each step of its iteration. Due to the way LARS algorithm is defined, this helps in understanding which predictors have the most correlation with the response, and give an idea about the tradeoff between model complexity (lesser variables) and better fit (more variables, lesser loss).
The algorithm is simple:

1. Center everything, normalize the predictors.
2. Initialize the solution to 0, an **active set** to an empty set, and a **sign vector** to 0.
3. Since initial solution is 0, initial residual $r = y - 0 = y$
4. Consider the predictor whose Pearson correlation with r is the largest (wlog X_1). Include it in the active set. Update the sign vector with the sign of the thus obtained correlation. Move your solution vector in this direction, ie $\beta = \gamma X_1$, where γ is slowly changed.
5. But how much do we move? Compute r and its correlation with the non active predictors as a function of γ, and analytically solve for it to get some other predictor having same correlation with the residual (in absolute value). (If you are interested, the formula is given by $\gamma = \min^{+}_{j \notin A}(\frac{C-c_j}{a_j-a}, \frac{C+c_j}{a_j+a})$, where C is the absolute value of the maximum correlation across all variables, a_j is the correlation of the non active predictor under consideration, and a is the common correlation of the active predictors with the resdiual. Why common? We will see in the next step.)
6. Include this (or these) new predictors in your active set. Update the sign vector with its (their) sign of correlation. You now want to move the **coefficients of active predictors** in a direction which leads to **equal change in correlation (and hence same correlation as it got to the active set due to same correlation)** between the predictors in the active set, and the residuals (equal in magnitude, and remains same in sign); if d is that direction, then $\beta' = \beta + \gamma d$, for some γ; and we want $x_a^T r' - x_a^T r = \alpha s_j \implies -\gamma X_A^T X d = \alpha s$. Now $Xd = X_A d$ as d has zero coordinates for non active predictors. Thus, the direction $d = (X_A^T X_A)^{-1} s$ (upto a scale), where s is the sign vector. Corresponding movement of residuals is in the direction $X_A d = u$, $u = X_A (X_A^T X_A)^{-1} s_A$. Since $u = r' - r$, all active predictors have equal correlation in magnitude with it; it is the **equiangular** direction of movement of the **residuals**.
7. Now compute the amount of movement γ again such that some non active predictor catches up in correlation with the residuals.
8. Keep on repeating: eventually we obtain the least squares solution.

But how is this "evolution of coefficients" obtained? This is represented by a graph called the "solution path". The solution path has for each predictor, on the y axis, the value of the estimated regression coefficient. These values should be plotted as a function of something which indicates how far we have moved from the initial solution (which is 0); this is captured by **arc length**, ie the length we have actually moved. If the current estimate is β and the new estimate is β', then $\beta' = \beta + \gamma d$, hence the distance moved is $\gamma||d||_1$. Note that instead of using the analytically calculated optimal γ, we can also compute this as a function of $\gamma \in [0, \gamma_{opt}]$. If we have changed the active set K times, then we can sum up the lengths of these and add the new length as a function of γ. For each of these lengths, we can compute the value of estimated coefficient for each predictor and plot it.

However the actual utility of LARS lies in the following modification: while choosing amount of movement γ at each step, you choose the first point where **either** a non active predictor has obtained equal correlation, **or** an active predictor has now coefficient zero. If the latter happens before the former, drop it from the active set, and continue similarly. This method is called **LARS-LASSO** or LARS with LASSO modfication, and the tremendous advantage is that it has the same solution path as LASSO (LASSO solution path means for a fixed tuning parameter, compute estimates, plot L1 norm of entire parameter on x axis and estimated regression coefficient for each parameter on y. We plot against L1 norm as it is piecewise linear in it). Hence instead of repeatedly computing LASSO for several tuning parameters and comparing, one obtains the entire path and can easily compare different sets of parameters. This is the efficiency of LARS-LASSO over standard LASSO.

A short note on Numerical Delta Method (needed for Rupsa's paper)

Suppose we have a parameter θ and its estimate $\hat{\theta}$ such that $r_n(\hat{\theta} - \theta) \xrightarrow{D} G$. Consider a function $g : \mathbb{R}^k \to \mathbb{R}$. We are interested the asymptotic distribution of $r_n(g(\hat{\theta}) - g(\theta))$ when g is directionally differentiable at θ $\forall h$. In particular, if $g'_\theta(h)$ is the directional derivative, then Shapiro's result says

$$r_n(g(\hat{\theta}) - g(\theta)) \xrightarrow{D} g'_\theta(G)$$

But this limiting distribution involves the parameter θ; hence we must find a way to estimate this distribution.

Fang and Santos propose the following: define $Z_n^* = r_n(\hat{\theta}^* - \hat{\theta})$, where * represents bootstrapped statistic. If $Z_n^*|data \xrightarrow{D} G$, then

$$g'_{\hat{\theta}}(Z_n^*) \xrightarrow{D} g'_\theta(G)$$

However, this needs us to know the analytical expression of the directional derivative, or atleast a consistent estimator for it. Hence Hong and Li proposed the following: if $\Delta_n^* = \frac{g(\hat{\theta} + \epsilon_n Z_n^*) - g(\hat{\theta})}{\epsilon_n}$, then if $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} G$, g is directionally differentiable at θ, $Z_n^*|data \xrightarrow{P} G$, $\epsilon_n \to 0, \sqrt{n}\epsilon_n \to \infty$, then

$$\Delta_n^*|data \xrightarrow{p} g'_\theta(G)$$

Hence the conditional ecdf of Δ_n^* consistently estimates $g'_\theta(G)$, and can be used to compute its quantiles.

Presentation Highlights

1 Speaker : Rupsa Ray (BS2225) - Kink Regression

Drive: https://drive.google.com/drive/folders/1sYYQGiPo2lN0kp5XYIqqxXVoxRug5zFa?usp=drive_link

1. The goal is to model a situation where slope of mean response for a covariate might change beyond a certain threshold.
2. Model: $y_t = \beta_1(x_t - \gamma)_- + \beta_2(x_t - \gamma)_+ + \beta_3' z_t + e_t$. Estimation is done by fixing various values of γ and then doing standard least squares and comparing the loss. The one yielding minimum loss is taken and then used to compute the least squares estimate for β.
3. Testing for threshold effect ie $\beta_1 = \beta_2$: Since the variables are generally time series, under some weak dependency conditions, moment conditions, the asymptotic distribution of the goodness of fit statistic ($n\frac{SSE_{H_0} - SSE_{full}}{SSE_{full}}$) is the supremum of a mean zero Gaussian process (stochastic process) with a certain covariance kernel. This is non standard so multiplier bootstrap is used; iid standard normals are multiplied with residuals to get new y_t^* and then the test statistic is computed for this data.
4. For testing $\gamma = \gamma_0$, gof statistic is asymptotically chi-squared with 1 dof. A wild bootstrap improves accuracy: multiply residuals with normal rvs to get new residuals, use estimated coefficients as surrogates for true coefficients to generate new data and repeatedly compute test statistic.
5. Confidence band for kink regression function: using numerical delta method (as the regression function is not differentiable).

2 Speaker : Saptashwa Baisya (BS2226) - Categorical ANOVA

Drive: https://drive.google.com/drive/folders/1l7p9gq_KLVw9SsIKjz9MqP8MjdYShf69?usp=drive_link

1. Setting: G many treatments, n subjects. Each subject receives one treatment and shows a response which falls in one of I categories. n_{ij} people receive jth treatment and show ith category of response. $n_{.j} = \sum_{i=1}^{I} n_{ij}$.
2. A multinomial likelihood is fit for each treatment: $L_j = \binom{n_{.j}}{n_{1,j},\ldots,n_{I,j}} \prod_{i=1}^{I} p_{ij}^{n_{ij}}$. Total likelihood $= \prod_{j=1}^{G} L_j$.
3. Variance for categorical rv with pmf $(p_i)_{i=1}^m$ is defined as $\frac{1}{2}(1 - \sum_{i=1}^m p_i^2)$. Hence total sum of squares of sample is $\frac{n}{2}(1 - \sum_{i=1}^{I}(\frac{n_{i.}}{n})^2)$.
4. Within treatment sum of squares is given as $WSS = \sum_{j=1}^{G}\left(\frac{n_{.j}}{2}(1 - \sum_{i=1}^{I}(\frac{n_{ij}}{n_{.j}})^2)\right)$. Between treatment SS is $BSS = TSS - WSS$. Non negativity of BSS can be shown by CS inequality (Titu's form).
5. Result: Asymptotically if $n_{.j} \to \infty \forall j$, then TSS and BSS are independent under $H_0 : p_{ij} = p_i \forall i, j$. Hence $C = \frac{(n-1)(I-1)BSS}{TSS}$ is approximately $\chi^2_{(I-1)(G-1)}$ under H_0 and can be used to test.
6. Why not use Fisher's exact or Chi-squared test? We want a decomposition of variance and an idea of what proportion of variability is explained by the model.

3 Speaker : Shreetama Bhuniya (BS2229) - Zero Inflated Poisson Regression

Drive: https://drive.google.com/drive/folders/1rXusoyj09zxBLAaoYX1XvQyOd2qaPSjx?usp=drive_link

1. Standard poisson regression model: $Y_i \sim poi(\exp(x_i^T\beta))$. Estimation done using MLE and Newton-Rhapson. However poisson has property that mean and variance is same; in practice it is often not the case. The paper focuses on the case when counts of zero are much more than what expects under poisson model.

2. Zero inflated poisson model is $Y_i \overset{D}{=} \theta_i(z_i)A+(1-\theta_i(z_i))B$, where $A \overset{a.e.}{=} 0, B \sim poi(\exp(x_i^T\beta))$, and z_i is a vector of covariates used to define θ_i in the following way: $\theta_i(z_i) = \text{logit}(z_i^T\gamma)$. Hence the entire parameter vector is (β, γ).

3. The log likelihood is maximized using EM algorithm as it appears diffciult to maximise.

$$\begin{aligned} l(\gamma, \beta, Y) &= \sum_{i=1}^{n} D_i \ln(\exp(z_i^T\gamma) + \exp(-\exp(x_i^T\beta))) \\ &+ \sum_{i=1}^{n}(1-D_i)(y_i x_i^T\beta - \exp(x_i^T\beta) - \ln y_i!) - \sum_{i=1}^{n}\ln(1+\exp(z_i)) \end{aligned}$$

where $D_i = \mathbb{I}(Y_i = 0)$. For using EM, missing values are introduced as $\delta_i = \mathbb{I}(\text{ith observation is from state 0})$. On introduction of δ_i, the log likelihood seperates into sum of two functions: one in β and one in γ. These can be maximised seperately.

$$l(\beta, \gamma, \delta, Y) = \sum_{i=1}^{n}[\delta_i z_i^T\gamma - \ln(1+\exp(z_i^T\gamma))] + \sum_{i=1}^{n}(1-\delta_i)[y_i x_i^T\beta - \exp(x_i^T\beta)] + const.$$

4. E step: Conditional expectation of δ_i is $(1+\exp(z_i^T\gamma - \exp(x_i^T\beta)))^{-1}\mathbb{I}(Y_i > 0)$.
M step: For β, a weighted log linear poisson regression is fit of Y on X with weight $1-\delta_k$. For γ, a logistic regression of Y on Z with weights $1-\delta_k$ is fit. (these fittings are equivalent to maximising likelihood).

5. Initial guesses: For β, the MLE of positive poisson log likelihood is suggested (ie poisson regression with 0s dropped). For γ, set all terms other than intercept to 0 and use log odds of excess probability of zero $(\frac{\#y_i=0-\sum_{i=1}^{n}\exp(-\exp(x_i^T\beta_0))}{n})$ as estimate of intercept.

6. Model comparison: Use a discrete randomized probability integral transform and compare histogram with uniform.

$$V = \begin{cases} F(Y-1) + U(F(Y) - F(Y-1)), & \text{when } Y \geq 1 \\ UF(0), & \text{when Y=0} \end{cases}$$

4 Speaker : Soumil Sangrajhka (BS2233) - COM Poisson Regression

Drive: https://drive.google.com/drive/folders/1HIIUDkGojd47CsQR_xTFEd9KSDPftLxn?usp=drive_link

1. Problem: To find a general model which can accommodate both under and over dispersed data. Poisson regression cannot handle over dispersion, negative binomial (poisson with gamma prior) cannot handle under dispersion.

2. COM (Conway Maxwell) Poisson distribution: $\mathbb{P}(Y = y_i) \propto \frac{\lambda^{y_i}}{(y_i!)^\nu}$. This includes poisson $(\nu = 1)$, geometric $(\nu = 0, \lambda < 1)$ and bernoulli $(\nu \to \infty)$. If the normalizing constant is called $Z(\lambda, \nu) = \sum_{i=1}^{\infty}\frac{\lambda^i}{(i!)^\nu}$, then $\mathbb{E}(Y) = \frac{\partial \ln(Z)}{\partial \ln \lambda}, \mathbb{V}(Y) = \frac{\partial \mathbb{E}(Y)}{\partial \ln \lambda}$. An approximate expression for mean is $\mathbb{E}(Y) \approx \lambda^{\frac{1}{\nu}} - \frac{\nu-1}{2\nu}$; from here one can get $\mathbb{V}(Y) = \frac{1}{\nu}\mathbb{E}(Y)$.

3. COM Poisson model: $Y_i|X_i \sim COM - Poisson(\exp(\beta^T x_i), \nu)$. Maximum likelihood estimation is done by Fishers scoring.

4. Obtaining fitted values (mean/median response): if $\hat{\nu} \leq 1$ or $\hat{\lambda}_i > 10^\nu$, then use the approximation for expectation. Else use the median response from inverse cdf (first point where the probability crosses 0.5).

5. Testing for dispersion (Poisson vs COM-Poisson): $H_0 : \nu = 1$. Likelihood ratio test can be used, for large sample it has asymptotic χ_1^2, for small sample use bootstrap.

6. Testing for significance of β_i: use asymptotic normality of MLE, and the fact that asymptotic dispersion is inverse of Fisher information. For small sample use bootstrap for the same statistic.

5 Speaker : Sourav De (BS2236) - Beta Regression for rates, proportions

Drive: https://drive.google.com/drive/folders/19aSIdBF4vx2czTFklt9YhXiDMDU4V5_n?usp=drive_link

1. For responses in the range (0,1), the beta distribution can be used to model, as it captures a wide variety of distribution shapes.

2. Model: $Y_t \sim Beta(\mu_t, \phi)$, where $\mathbb{E}(Y_t) = \mu, \mathbb{V}(Y) = \frac{\mu_t(1-\mu_t)}{1+\phi}$. ϕ is a precision parameter in the sense that large values of it means lesser variance for fixed mean.
$g(\mu_t) = x_t^T\beta$, where g is a strictly monotone link function, which can be differentiated twice, mapping the unit interval to the entire real line. Note that heteroscedasticity is accommodated in the model in a certain sense.
Some common link functions are logit, probit, log-log.

3. Using logit link leads to a nice interpretation. If the ith covariate is increased by c units, then $\exp(c\beta_i) = \frac{\frac{\mu'}{1-\mu'}}{\frac{\mu}{1-\mu}}$, where μ' is the new mean.

4. Log likelihood does not admit closed form solutions. For numerical methods, initial guess for β is suggested to be the least square estimator for $g(Y)$ on X. A delta method based guess for ϕ is $\frac{1}{n}\sum_{t=1}^{n}\frac{\tilde{\mu_t}(1-\tilde{\mu_t})}{MSE}g'(\tilde{\mu_t})^2 - 1$, where $\tilde{\mu_t} = g^{-1}(x_t\beta_0)$.

5. For checking significance of individual coefficients, log likelihood ratio can be used; it converges in law to $\chi^2_{dim(\beta)}$, whenever a unique maximiser exists which is CAN.

6. Walds test can also be used: test statistic $(\hat{\beta}-\beta_0)^T I_{\hat{\beta}}^{-1}(\hat{\beta}-\beta_0)$, where $I_{\hat{\beta}}$ is the submatrix of the Fisher information, for β evaluated at $\hat{\beta}$. It is also under H_0 aymptotically $\chi^2_{dim(\beta)}$.

6 Speaker : Souvik Roy (BS2237) - Propensity Score Weighting

Drive: https://drive.google.com/drive/folders/1ra1nHdd4fgCiXlDQ9D_DmX5GW32qDDy5?usp=drive_link

1. Problem: When we want to compare two groups, one who received treatment and one who did not, propensity score weighting method is used to reweight people on the basis of how likely they were to receive the treatment, given their characteristics. This is ultimately used for estimation of some parameter.

2. Suppose we have H many clusters, and a sample of size $n_h, h = 1(1)H$ is drawn from each, units indexed by $k = 1(1)n_h$; $n = \sum n_h$. Z_{hk} is the indicator for treatment or control. $X_{hk} = (U_{hk}, V_h)$ are covariates - unit and cluster level. Response Y_{hk} is observed. The propensity score is defined as $e(X) = \mathbb{P}(Z = 1|X)$.

3. The paper focuses on unmeasured cluster level confoundedness (ie effect of some cluster level factors, ie how good is the hospital, etc), hence V_h is absent in all models.

4. Descriptive comparison - We want to compare responses of two groups having uncontrollable covariates, eg. Race, Sex, Birth Year, etc. We look at disparities across the levels of such a covariate (with balanced distribution of other covariates), but we cannot comment on the causality.
Causal Comparison - We want to assess effect of a treatment whose assignment is under our control. We want to estimate the causal effect; what would happen to the same person if he got the alternate treatment (in reality only one of the two treatments is observed per person).

5. For descriptive we want to estimate population average controlled difference;
$\pi_{ACD} = \mathbb{E}_X[\mathbb{E}(Y|X = x, Z = 1) - \mathbb{E}(Y|X = x, Z = 0)]$
For causal, under SUTVA assumption (my treatment does not affect your outcome; each treatment is well defined and consistently applied), $Y_{hk}(z)$ has two potential outcomes based on $z = 0/1$: $Y_{hk}(Z) = Y_{hk}(1)Z_{hk} + Y_{hk}(0)(1 - Z_{hk})$ We want to estimate average treatment effect: $\pi_{ATE} = \mathbb{E}(Y(1) - Y(0))$. Estimation of causal effect from observational data is performed by assuming $Y(1), Y(0) \perp Z|X$ (unconfoundedness). Under this assumption, $\pi_{ATE} = \pi_{ACD}$.

6. Under unconfoundedness, $\mathbb{E}\left(\frac{ZY}{e(X)} - \frac{1-ZY}{1-e(X)}\right) = \pi_{ACD} = \pi_{ATE} = \pi$.

7. Models for propensity score:

   (a) When many small clusters, and we believe all confounding (influence by any other factor) is explainable by the observed covariates:

   $$\text{logit } e_{hk} = \delta_0 + \alpha^T X_{hk}$$

   If clusters differ in unobserved ways, there will be bias in this.

   (b) When moderate/large clusters, and suspect important cluster effect, then:

   $$\text{logit } e_{hk} = \alpha^T U_{hk} + \delta_h$$

   Needs both treatment and control in each cluster.

   (c) When many small/sparse clusters and some clusters dont have both in them:

   $$\text{logit } e_{hk} = \delta_h + \alpha^T X_{hk}, \delta_h \sim N(\delta_0, \sigma_\delta^2)$$

8. Estimators (SE via bootstrap):

   (a) Non parametric marginal estimator (ignores clustering):

   $$\hat{\pi}^{\text{ma}} = \frac{\sum_{Z_{hk}=1} Y_{hk} w_{hk}}{w_1} - \frac{\sum_{Z_{hk}=0} Y_{hk} w_{hk}}{w_0},$$

   where w_{hk} is the inverse-probability weight of subject k in cluster h, based on the estimated propensity score (e.g., from one of the three models before), with:

   $$w_{hk} = \begin{cases} \frac{1}{\hat{e}_{hk}} & \text{if } Z_{hk} = 1 \\ \frac{1}{1-\hat{e}_{hk}} & \text{if } Z_{hk} = 0 \end{cases}$$

   and

   $$w_z = \sum_{h,k:Z_{hk}=z} w_{hk} \quad \text{for } z = 0, 1.$$

   (b) Non parametric clustered estimator (first computes ACD for each cluster):

   $$\hat{\pi}_h = \frac{\sum_{k\in h:Z_{hk}=1} Y_{hk} w_{hk}}{w_{h1}} - \frac{\sum_{k\in h:Z_{hk}=0} Y_{hk} w_{hk}}{w_{h0}},$$

   where $w_{hz} = \sum_{k\in h:Z_{hk}=z} w_{hk}$ for $z = 0, 1$, and then takes their mean weighted by the total weights in each cluster, $w_h = \sum_{k\in h} w_{hk}$:

   $$\hat{\pi}^{\text{cl}} = \frac{\sum_h w_h \hat{\pi}_h}{\sum_h w_h}.$$

   (c) Parametric doubly robust estimator (robust to correct choice of propensity model or potential outcome model (will see next)):

   $$\hat{\pi}^{\text{dr}} = \frac{1}{n} \sum_{h,k} \hat{\pi}_{hk}, \tag{9}$$

   where

   $$\hat{\pi}_{hk} = \left[\frac{Z_{hk} Y_{hk}}{\hat{e}_{hk}} - \frac{(Z_{hk} - \hat{e}_{hk})\hat{Y}^1_{hk}}{\hat{e}_{hk}}\right] - \left[\frac{(1 - Z_{hk}) Y_{hk}}{1 - \hat{e}_{hk}} + \frac{(Z_{hk} - \hat{e}_{hk})\hat{Y}^0_{hk}}{1 - \hat{e}_{hk}}\right],$$

   with $\hat{Y}^z_{hk}$ being the fitted (potential) outcome from an outcome model in group z.

9. Models for potential outcome (analogous to propensity score models):

   (a) Marginal: $Y_{hk} = \eta_0 + Z_{hk}\gamma + X_{hk}\beta + \epsilon_{hk}$

   (b) Fixed effects: $Y_{hk} = \eta_h + Z_{hk}\gamma + U_{hk}\beta + \epsilon_{hk}$

   (c) Random effects: Same as fixed, use normal distribution prior for η_h.

10. A large sample property: Under a specific data generating mechanism which violates standard propensity score assumptions, ignoring the clustered structure (ie marginal model) leads to biased estimates; accounting for them gives consistency.

7 Speaker : Supratim Das (BS2243) - Efficient algorithm for L1 logistic regression

Drive: https://drive.google.com/drive/folders/1oMrLMSqjnxVuzF0wt0RHR8xv_XrSyneI?usp=drive_link

1. The basic model is the standard logistic regression model. However instead of getting MLE, we maximise the likelihood under a constraint.

$$\text{maximise } l(\theta|x_1, y_1, \dots) = \sum_{i=1}^{n} \ln p(y_i|x_i, \theta), \text{ subject to } ||\theta||_1 \leq C$$

where $p(1|x,\theta) = \frac{1}{1+\exp(-x^T\theta)}$, $p(0|x,\theta) = 1 - p(1|x,\theta)$.

2. In the unconstrained cases, a **damped Newton Rhapson** may be used to get MLE: ie

$$\begin{aligned} \gamma^{(k)} &= \theta^{(k)} - H^{-1}(\theta^{(k)})\nabla(\theta^{(k)}) \\ \theta^{(k+1)} &= (1-t)\theta^{(k)} + t\gamma^{(k)} \end{aligned}$$

t tries to control overshooting in the wrong direction, and is found using a line search (can be exact: optimize the function over t, which is generally difficult, or heuristics: decrease t until a sufficient decrease in function value).

3. The paper shows that $\gamma^{(k)}$ can be computed by solving a least squares at each step, specifically,

$$\gamma^{(k)} = argmin_\gamma(||\Lambda^{0.5}X^T\gamma - \Lambda^{0.5}z||_2^2), \Lambda = diag\left(\frac{\exp(x_i^T\theta^{(k)})}{(\exp(x_i^T\theta^{(k)})+1)^2}\right)$$

$$z_i = x_i^T\theta^{(k)} + \frac{\exp(x_i^T\theta^{(k)})}{1+\exp(x_i^T\theta^{(k)})}\frac{y_i}{\Lambda_{ii}}$$

This is an IRLS procedure in the sense that at each step the design matrix is same but the "response" and weights change.

4. For regularised case, an estimate is obtained by solving the penalised version of the least squares problem at each step. Penalised regression can be solved by using Least Angle Regression with LASSO modification.

5. The algorithm proposed by the paper solves the penalised regression at each step using LARS and then gets $\theta^{(k+1)}$. The algorithm converges to the global optimum of the penalised logistic regression if we start from $\theta^{(0)} = 0$.

8 Speaker : Swagato Das (BS2245) - Factor selection in ANOVA

Drive: https://drive.google.com/drive/folders/1TdGUDls7_8HzEZQfsQEtyVfDGzB9NqTY?usp=drive_link

1. Problem: want to do factor selection in an ANOVA like setting, but not necessarily in balanced design. Using methods like LASSO selects more factors than usual as it deals with the regression set-up and selects individual factor level variables instead of factors as a whole. Also, if wlog we consider centered response and covariates in the model $Y = \sum_{j=1}^{J} X_j\beta_j + \epsilon, [X_j]_{n\times p_j}$, J being number of factors, and wlog X_j are orthonormal (appropriately reparametrize the model), then LASSO selection is dependent on the reparametrization chosen.

2. Method 1 (Group LASSO): given positive definite matrices $K_1, \dots, K_J$, the group lasso estimate is defined as:

$$\hat{\beta} = argmin(\frac{1}{2}||Y - \sum_{j=1}^{J} X_j\beta_j||^2 + \lambda\sum_{j=1}^{J}||\beta_j||_{K_j}), ||\beta_j||_{K_j} = \sqrt{\beta_j^T K_j \beta_j}$$

Common choices are $K_j = I_{p_j}$ or $p_j I_{p_j}$. For these choices, the factor selection is independent of the choice of parameterization.

3. Proposition: $(\beta_i)_{i=1}^J$ is a solution iff $X_j^T(Y - X\beta) = \lambda\sqrt{p_j}\frac{\beta_j}{||\beta_j||}$ for all non zero β_j and $||X_j(Y - X\beta)|| \leq \lambda\sqrt{p_j}$ whenever $\beta_j = 0$. This yields the following iterative equation: $\beta_j^{(k+1)} = (1 - \lambda\frac{\sqrt{p_j}}{||S_j||})_+ S_j$ where $S_j = X_j^T(Y - X\beta_{-j}^{(k)})$. Convergence is stable.

4. Method 2 (Group LARS): A natural extension of LARS here is by using the angle made by residuals with the column space of a particular factor instead of highest correlation (basically angle) with the coded variables. For a vector v and projector P, the angle is given by $\cos^2\theta = \frac{||Pv||_2^2}{||v||_2^2}$. The standard extension is thus to consider the factor with smallest angle (analogue to highest correlation), or largest $\cos^2$ of angle, (possibly standardized by number of levels if they are unequal) and include it in the active set, then move in a direction given by $\gamma_A = (X_A^T X_A)^{-1} X_A r$, where r is the residual; since least squares projection is the unique vector making equal angle with all covariates. Then we compute how far we can move before some inactive predictor becomes as correlated as the covariates in the active set, and update $\beta' = \beta + \alpha\gamma_A$. This yields the entire solution path; since we have grouped data we plot $||\beta_i||_1$ against arc length for all groups. Using this path we can do factor selection.

5. Method 3 (Group Non Negative Garrotte): Non Negative Garrotte is a variable selection technique which is done in the following way: compute LSE and then compute $d(\lambda) = argmin_d \left(\frac{1}{2}||(Y - Zd)||^2 + \lambda\sum_{j=1}^J d_j\right), Z_{n\times p} = [X_1\hat{\beta}_1, \ldots, X]$. The NNG estimate is given by $d_j(\lambda)\hat{\beta}_{j\,LS}$. Its advantage is that it is theoretically better than LASSO as it is path consistent; for atleast one λ the estimates are consistent as well consistent in variable selection; however practical usage is very limited. Here we have p_j levels for each factor, hence a natural extension is to use the same scaling for the coded variables for each factor. Solution path is constructed using a similar LARS-LASSO algorithm.

6. Tuning the models: For Gaussian regression problems, a ue for the true risk $(\mathbb{E}(\frac{||\hat{\mu}-\mu||^2}{\sigma^2}))$ is $C_p(\hat{\mu}) = \frac{||\hat{\mu}-Y||^2}{\sigma^2} - n + 2df_{\mu,\sigma^2}, df = \sum_{i=1}^n \frac{cov(Y_i,\hat{\mu_i})}{\sigma^2}$ which is estimated generally using bootstrap. Here, the authors define unbiased estimates (for orthonormal designs) for the df term for each of the 3 methods. We prefer the one with smaller C_p.

9 Speaker : Tanishq Prasad (BS2247) - FDR and BH method

Drive: https://drive.google.com/drive/folders/1-QWxTxs9HZ-fBAZwnHfAOiK8ezj4hrhj?usp=drive_link

1. The paper gives a new philosophy for error control in multiple comparisons and a method to control it.

2. Familywise error rate is defined as $\mathbb{P}(V \geq 1)$ where V is number of false rejections. Strong control of FWER (in all configurations) greatly reduces power.

3. False discovery rate is defined as $\mathbb{E}(Q_e)$ where $Q_e = \frac{V}{R}, R > 0$ and 0 otherwise- V being number of false nulls rejected and R being total number of rejections. This expectation is computed under a particular configuration of which nulls are true and which are not.

4. $Q_e \leq FWER$, hence it is a more relaxed criteria in hopes of getting more power.

5. Procedure: Order the p-values, start from the largest. The moment $p_{(i)} \leq \frac{i}{m}q^*$, reject all nulls with p values $\leq p_{(i)}$, and accept the rest. It controls FDR for independent p values under all configurations at level q^*.

6. Later in 2001 it was shown that a harmonic correction would control FDR irrespective of independence (ie dividing the original threshold by $\sum_{j=1}^i \frac{1}{j}$).

10 Speaker : Urjit Paul Chowdhury (BS2248) - HIV Classification using logistic regression

Drive: https://drive.google.com/drive/folders/1GGPXHuph8HZ_BCtx-Z2hjlvwWJuYMG2N?usp=drive_link

1. Problem: A main dataset having some covariates $\mathbf{X} = (X_1, \ldots, X_p)$ needs to be used to predict whether a person has HIV or not, which is an unobserved response. The model which one wishes to use is $\mathbb{P}(Y_i = 1) = \text{ logit}^{-1}(\beta_0 + x_i^T\beta)$, the standard logistic regression. But responses are unobserved so we rely on our knowledge from previous studies to solve the problem.

2. We have a contingency table (from previous studies) for a subset of predictors, $T = (X_1, \ldots, X_r)$, where their space is divided into K disjoint cells $C_1, \ldots, C_K$. Each cell has observed counts for $m_k^{(1)}, m_k^{(0)}$ for $Y = 1, 0$. The likelihood is:

$$L = \prod_{h=0}^{1} \prod_{k=1}^{K} \mathbb{P}(Y = h, T \in C_k)^{m_k^{(h)}}$$

3. Assumptions:

   (a) For both datasets, the logistic model is true, except for possibly β_0.

   (b) $\mathbb{P}(\mathbf{X}|T \in C_K)$ is same in both datasets.

4. Conditional distribution of $\mathbf{X}$ given $T \in C_K$ is estimated from main data empirically: $\hat{\mathbb{P}}_{\mathbf{X}}(x|T \in C_K) = \frac{\sum_{j:\mathbf{x}_j = x} \mathbb{I}(T_j \in C_K) w_j}{\sum_{j:\mathbf{x}_j = x} w_j}$, where w_j are sampling weights. Hence,

$$\hat{\mathbb{P}}(Y = h|T \in C_K) = \sum_{i:T_i \in C_K} \mathbb{P}(Y = h|\mathbf{X} = x, T \in C_K)\hat{\mathbb{P}}_{\mathbf{X}}(x|T \in C_K)$$

   (also note $\{\mathbf{X} = x\}$ is a subset of $\{T \in C_K\}$ whenever the intersection is non empty). Since $\mathbb{P}(T \in C_K)$ is independent of parameters of interest, we work with conditional likelihood and maximise it numerically to get estimates.

5. The data from previous studies may oversample cases where HIV is present since people generally go when they face symptoms. Hence β_0 is calibrated to meet some population prevelance estimator: fix other parameters at MLE and then solve:

$$\sum_{i=1}^{n} \mathbb{P}\big(Y_i = 1 \mid X_i; \hat{\beta}\big) \times w_i \;=\; \hat{\hat{\mathbb{P}}}(Y = 1)$$

   , with $\hat{\hat{\mathbb{P}}}$ is some previously known epidemiological estimator for prevelance.

6. Under identifiability, finite second moment of predictors and $\mathbb{P}(Y = h, T \in C_k) > 0$, $\hat{\beta}$ converges to β in probability and has asymptotic normality (both m,n go to infinity; the main data size and contingency table size). A bounded in probability order is also mentioned for calibrated β_0.