Elliptical Predictors for Logistic Regression

Will Potts
29 August 2007
Risk Management, Capital One

- Benefits of coordinate-wise normalizing transformations
  - Link-free consistency
    In an elliptically-contoured predictor space, simple models are closer to the true model.
  - Reduce nonlinear confounding
    Nonlinear confounding can result in inferior prediction due to model misspecification.
  - Local adaptivity
    Suitable transformations allow models to balance their sensitivity to dense and sparse regions of the predictor space.
- The practice of transforming the predictors
  - Fitting Johnson's S_U family of transformations
  - Alternatives

Link-free consistency

— Li and Duan (1989). Regression analysis under link violation. *The Annals of Statistics*.

1D regression: $y \perp\!\!\!\perp \mathbf{x} \,\big|\, \mathbf{x}\boldsymbol{\beta}$ where $\mathbf{x}\boldsymbol{\beta} = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$

$$\Rightarrow \quad \ln\left(\frac{E(y \mid \mathbf{x})}{1 - E(y \mid \mathbf{x})}\right) = \beta_0 + \ln\left(\frac{f(\mathbf{x}\boldsymbol{\beta} \mid y=1)}{f(\mathbf{x}\boldsymbol{\beta} \mid y=0)}\right) = \beta_0 + \overbrace{g(\mathbf{x}\boldsymbol{\beta})}^{\substack{\text{unknown (nonlinear)} \\ \text{function}}}$$

Linear design condition: $E(\mathbf{x} \mid \mathbf{x}\boldsymbol{\beta}) = E(\mathbf{x} \mid \mathbf{x}\boldsymbol{\beta}, y)$ is linear

$$\Leftarrow \quad \mathbf{x} \mid y \sim \underbrace{\text{elliptically contoured}}_{\substack{\text{e.g.} \\ \text{multvariate normal} \\ \text{multivariate } t}} \quad \Leftrightarrow \quad E(\mathbf{x} \mid \mathbf{x}\boldsymbol{\alpha}, y) \text{ is linear, for all } \boldsymbol{\alpha}$$

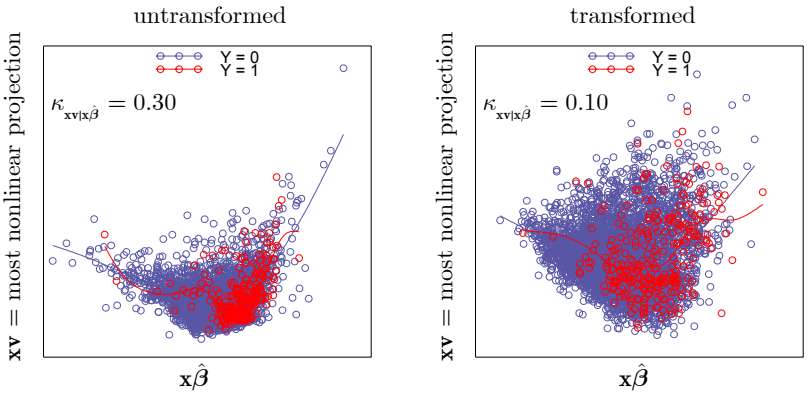
Consequent: The MLE of $\boldsymbol{\beta}$ from the logit-linear approximation is (Fisher) consistent.

$$\ln\left(\frac{E(y \mid \mathbf{x})}{1 - E(y \mid \mathbf{x})}\right) = \beta_0 + \mathbf{x}\boldsymbol{\beta}$$

Nonlinear confounding

"… relationship between y and $\mathbf{x}$ is confounded with the nonlinear relationship between the regressor variables. … We regard nonlinear confounding as an intrinsic weakness in the data."

— Li (1997). Nonlinear confounding in high-dimensional regression. *The Annals of Statistics*.



The kappa measure of nonlinear confounding

Nonlinear confounding obscures the relationship between the outcome and the predictors.

The proportion $\kappa_{\boldsymbol{\beta}}$ measures the strength of the nonlinear relationship between linear combinations of the predictors and $\mathbf{x}\boldsymbol{\beta}$.

$$\kappa_{\mathbf{xv}|\mathbf{x}\boldsymbol{\beta}} = \frac{\text{var}(k(\mathbf{x}\boldsymbol{\beta}))}{\text{var}(k(\mathbf{x}\boldsymbol{\beta}) + \varepsilon)} \quad \text{where } \mathbf{xv} = \underbrace{l(\mathbf{x}\boldsymbol{\beta})}_{\text{linear}} + \underbrace{k(\mathbf{x}\boldsymbol{\beta})}_{\text{nonlinear}} + \varepsilon$$

$$\kappa_{\boldsymbol{\beta}} = \max_{\rho(\mathbf{xv},\mathbf{x}\boldsymbol{\beta})=0} \left(\kappa_{\mathbf{xv}|\mathbf{x}\boldsymbol{\beta}}\right)$$

$$\arg\max_{\rho(\mathbf{xv},\mathbf{x}\boldsymbol{\beta})=0} \left(\kappa_{\mathbf{xv}|\mathbf{x}\boldsymbol{\beta}}\right) = \text{ most nonlinear projection}$$

Large values of $\kappa_{\boldsymbol{\beta}}$ indicate departures from the linear design condition.

$$\kappa_{\boldsymbol{\beta}} = 0 \quad \Leftrightarrow \quad E(\mathbf{x}|\mathbf{x}\boldsymbol{\beta}) = E(\mathbf{x}|\mathbf{x}\boldsymbol{\beta}, y) \text{ is linear}$$

Hence, large values of $\kappa_{\boldsymbol{\beta}}$ indicate that the linear approximation is deficient. Yet nonlinear confounding can make the linear approximation appear to fit better than it does (overlinearization). Nonlinear confounding also degrades the validity and power of the test H_0: $\boldsymbol{\beta} = 0$, based on the linear approximation.

Estimating kappa using SIR (sliced inverse regression)

Estimates of κ_β can be used to evaluate the effectiveness of the predictor transformations and as a general regression diagnostic.

```
proc logistic data=tr des;
model y=x1 x2 x3 x4;
output out=o xbeta=xbeta;
run;

proc glm data=o noprint;
model x1 x2 x3 x4=xbeta;
output out=o2 r=r1 r2 r3 r4;
quit;

proc rank data=o2 groups=50 out=s;
var xbeta;
ranks slice;
run;

proc candisc data=s out=v;
class slice;
var r1 r2 r3 r4;
run;
```

Forward linear regression. Treat the linear predictor (`xbeta`) as the SIR output variable.

Inverse regression. Treat the residuals as the SIR input variables.

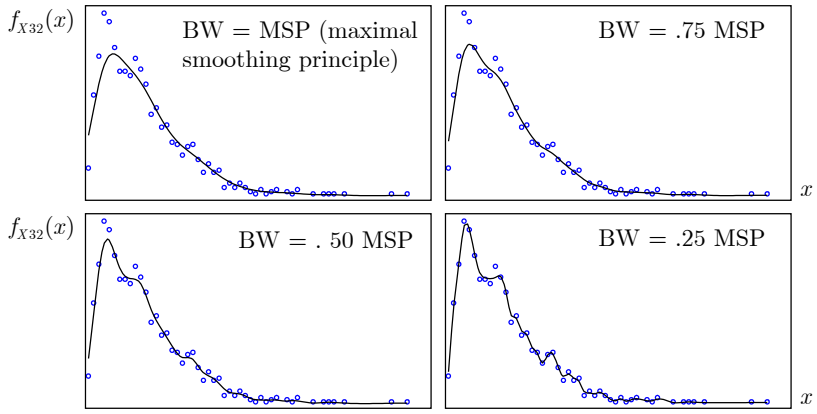
Sort and slice the output.

Eigenvalue decomposition. Kappa is the first squared canonical correlation. The most nonlinear projection is the first canonical variable.

- Li (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*.
- Cook (1998). *Regression Graphics*, Wiley.

Nonparametric density estimation

Lessons learned from nonparametric density estimation apply to flexible modeling the log-density ratio.



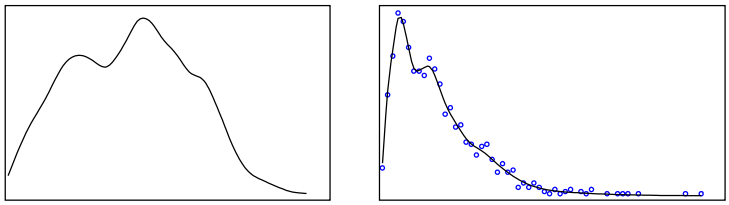
A global bandwidth that fits well where the data are sparse, often oversmooths in dense regions. A global bandwidth that fits well where the data are dense, often undersmooths in sparse regions.

Locally-adaptive smoothing via transformation

What is needed is a local bandwidth that adapts to the smoothing requirements in each region. A computational shortcut is transform the data to a scale that is appropriate for global smoothing then back-transform the results to the original scale.

$$f_X(x) = f_T(t)\left|\left(\frac{d(T^{-1}(t))}{dt}\right)^{-1}\right|$$

$f_T(t)$



$T(x) = t$ $\qquad$ $x = T^{-1}(t)$

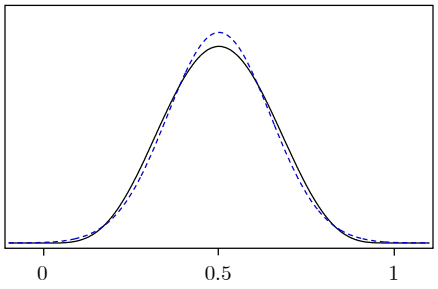
— Terrell (1990). The maximal smoothing principle in density estimation. *Journal of the American Statistical Association.*

— Wand, Marron, and Ruppert (1991). Transformations in density estimation (with discussion). *Journal of the American Statistical Association.*

— Yang and Marron (1999). Iterated transformation-kernel density estimation. *Journal of the American Statistical Association.*

Which transformation works best for global smoothing?

Answer: The transformation that makes the distribution closest to a Beta(5.3,5.3) distribution. This is the easiest density to estimate with a global bandwidth (with respect to mean integrated absolute error). The normal distribution is nearly as easy (ARE = 0.91).



Beta(5.3,5.3)
excess kurtosis = – 0.44

normal ------

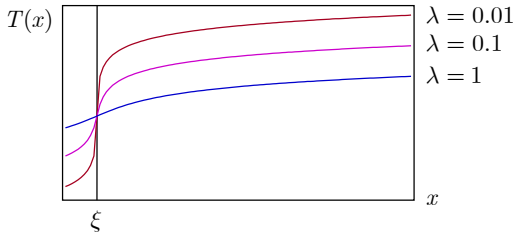
The lognormal is surprisingly difficult to estimate with a global bandwidth (ARE = 0.11). It is more difficult than many distributions with more skewness, more kurtosis, and more modes.

— Wand and Devroye (1993). How easy is a given density to estimate? *Computational Statistics and Data Analysis.*

Johnson's S_U family of transformations

$$T(x) = \gamma + \delta \sinh^{-1}\left(\frac{x-\xi}{\lambda}\right) = \gamma + \delta \ln\left(\frac{x-\xi}{\lambda} + \left(\left(\frac{x-\xi}{\lambda}\right)^2 + 1\right)^{1/2}\right)$$

The S_U family can normalize a variety of unbounded, unimodal distributions with varying amounts of skewness and kurtosis. It can pull-in long tails more effectively than the log transformation. Unlike the log, it can handle zeros and negative values.



- Johnson (1949). Systems of frequency curves generated by methods of translation. *Biometrika*.
- Burbidge, Magee, and Robb (1988). Alternative transformations to handle extreme values of the dependent variable. *Journal of the American Statistical Association.*

Selecting the optimal transformation

The optimal transformation can be estimated by nonlinear regression using the normal scores as the response variable. In practice, the parameter ξ can be set zero. The parameters γ and δ are needed for estimating the transformation but can be dropped when applying the transformation.

$$\Phi^{-1}\left(\frac{\text{rank}(x) - \frac{3}{8}}{n + \frac{1}{4}}\right) = \gamma + \delta \ln\left(\frac{x}{\lambda} + \left(\left(\frac{x}{\lambda}\right)^2 + 1\right)^{1/2}\right)$$

```
proc rank data=tr(where=(y=1) keep=y x)
          out=nb ties=high normal=blom;
var x;
ranks rx;
run;

proc nlin data=nb outest=oe save maxiter=50;
parms gam=-1,0 del=.5,1 lam=1,.1,.001;
model rx=gam+del*log(x/lam+sqrt((x/lam)**2+1));
quit;
```

- Lin and Vonesh (1989). An empirical nonlinear data-fitting approach for transforming data to normality. *The American Statistician.*