

General Formulation

Consider exponential family responses with

$$f(y|x) = \exp\{(y\theta(x) - b(\theta(x)))/a(\phi) + c(y, \phi)\},$$

where θ is the canonical parameter and $a(\phi)$ is the dispersion. Let η be a monotone transform of θ taking values on $(-\infty, \infty)$.

Observing $Y_i|x_i \sim f(y|x_i)$, one minimizes the PLK functional

$$-\frac{1}{n} \sum_{i=1}^n \{Y_i \theta(\eta(x_i)) - b_i(\theta(\eta(x_i)))\} + \frac{\lambda}{2} J(\eta)$$

for $\eta \in \mathcal{H} = \bigoplus_{\beta=0}^p \mathcal{H}_\beta$, where $J(f) = J(f, f) = \sum_{\beta=1}^p \theta_\beta^{-1}(f, f)_\beta$ and $(f, g)_\beta$ are IPs in \mathcal{H}_β with RKs kernels $R_\beta(x, y)$. The minimizer η_λ has an expression

$$\eta(x) = \sum_{\nu=1}^m d_\nu \phi_\nu(x) + \sum_{i=1}^n c_i R_J(x_i, x) = \boldsymbol{\phi}^T \mathbf{d} + \boldsymbol{\xi}^T \mathbf{c},$$

where $\{\phi_\nu\}$ is a basis of $\mathcal{N}_J = \mathcal{H}_0$ and $R_J(x, y) = \sum_{\beta=1}^p \theta_\beta R_\beta(x, y)$.

Slide 1

Iterated Weighted Least Squares

Taking derivatives of $l(\eta) = Y\theta(\eta) - b(\theta(\eta))$, one has

$$u(\eta) = \frac{dl}{d\eta} = -\left(y - \frac{db}{d\theta}\right) \frac{d\theta}{d\eta},$$

$$w(\eta) = \frac{d^2 l}{d\eta^2} = \frac{d^2 b}{d\theta^2} \left(\frac{d\theta}{d\eta}\right)^2 - \left(y - \frac{db}{d\theta}\right) \frac{d^2 \theta}{d\eta^2}.$$

The quadratic approximation of $Y_i \theta(\eta(x_i)) - b_i(\theta(\eta(x_i)))$ at $\tilde{\eta}$ is seen to be $(1/2)\tilde{w}_i(\tilde{Y}_i - \eta(x_i))^2 + C_i$, where $\tilde{Y}_i = \tilde{\eta}(x_i) - \tilde{u}_i/\tilde{w}_i$ and C_i is a constant. The PLK functional can thus be minimized iteratively through penalized weighted LS problems

$$-\frac{1}{n} \sum_{i=1}^n \tilde{w}_i (\tilde{Y}_i - \eta(x_i))^2 + \lambda J(\eta).$$

For fixed λ , this is the Newton iteration.

Slide 2

Smoothing Parameter Selection

Within an exponential family, the discrepancy between (η, ϕ) and (η_λ, ϕ) can be measured by the Kullback-Leibler distance

$$\begin{aligned} \text{KL}(\eta, \eta_\lambda) &= E_\eta[Y(\theta - \theta_\lambda) - (b(\theta) - b(\theta_\lambda))] \\ &= \{\mu(\theta - \theta_\lambda) - (b(\theta) - b(\theta_\lambda))\}, \end{aligned}$$

Slide 3

where $\mu = E[Y] = \dot{b}(\theta)$, or its symmetrized version,

$$\text{SKL}(\eta, \eta_\lambda) = \text{KL}(\eta, \eta_\lambda) + \text{KL}(\eta_\lambda, \eta) = (\mu - \mu_\lambda)(\theta - \theta_\lambda).$$

As an estimate of η based on data collected from $x_i, i = 1, \dots, n$, the performance of η_λ is to be assessed via the loss

$$L(\eta, \eta_\lambda) = \frac{1}{n} \sum_{i=1}^n (\mu - \mu_\lambda)(x_i)(\theta - \theta_\lambda)(x_i) \approx \frac{1}{n} \sum_{i=1}^n w(\eta'_i)(\eta - \eta_\lambda)^2(x_i).$$

Indirect Cross-Validation

It can be shown that $\tilde{Y}_i = \eta(x_i) + \tilde{\epsilon}_i + o_p(1)$, where $E[\tilde{\epsilon}_i] = 0$ and $E[\tilde{\epsilon}_i^2] = a(\phi)/w_i$. Fixing $\tilde{\eta}$, $U_w(\lambda)$ or $V_w(\lambda)$ select the minimizer $\eta_{\lambda, \tilde{\eta}}$ of the PWLS functional that approximately minimizes

$$\frac{1}{n} \sum_{i=1}^n w_i (\eta_{\lambda, \tilde{\eta}}(x_i) - \eta(x_i))^2,$$

Slide 4

which is a good proxy of $L(\eta, \eta_\lambda)$.

One may iterate on the PWLS problem, jointly updating $(\lambda, \tilde{\eta})$, with the $U_w(\lambda)/V_w(\lambda)$ -selected $\eta_{\lambda, \tilde{\eta}}$ as the new $\tilde{\eta}$.

- The iteration usually converges in 5-10 steps, but with no guarantee.
- When $a(\phi)$ is known, $U_w(\lambda)$ is usually preferred to $V_w(\lambda)$. $M_w(\lambda)$ can also be used, but with no Bayesian interpretation.

Direct Cross-Validation

For $\theta = \eta$, consider the relative Kullback-Leibler distance

$$\text{RKL}(\eta, \eta_\lambda) = n^{-1} \sum_{i=1}^n \{-\mu(x_i)\eta_\lambda(x_i) + b(\eta_\lambda(x_i))\}.$$

Replacing $\mu(x_i)\eta_\lambda(x_i)$ by $Y_i\eta_\lambda^{[i]}(x_i)$, one has

$$V_0(\lambda) = -\frac{1}{n} \sum_{i=1}^n \{Y_i\eta_\lambda(x_i) - b(\eta_\lambda(x_i))\} + \frac{1}{n} \sum_{i=1}^n Y_i(\eta_\lambda(x_i) - \eta_\lambda^{[i]}(x_i)).$$

Substituting $\eta_\lambda^{[i]}(x_i)$ by $\eta_{\lambda, \eta_\lambda}^{[i]}(x_i)$ and using the fact that

$$\eta_\lambda(x_i) - \eta_{\lambda, \eta_\lambda}^{[i]}(x_i) = \frac{a_{i,i}}{1 - a_{i,i}} \frac{Y_i - \mu_\lambda(x_i)}{\tilde{w}_i},$$

some hand-waving yields

$$V_g(\lambda) = L(\eta_\lambda | \text{data}) + \frac{\text{tr}(A_w W^{-1})}{n - \text{tr} A_w} \frac{1}{n} \sum_{i=1}^n Y_i(Y_i - \mu_\lambda(x_i)).$$

Slide 5

Direct CV, Empirical Performance

- The direct cross-validation score given above works well for the binomial and gamma families, but not so well for other families.

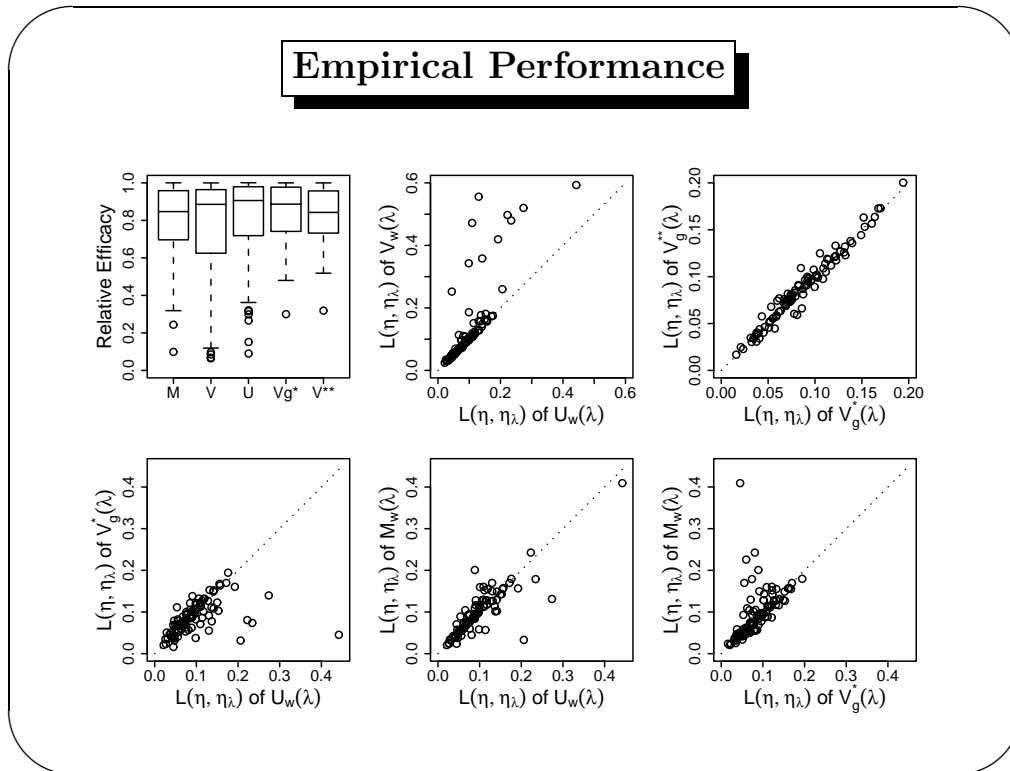
- Direct cross-validation is easier to work with for multiple smoothing parameters.

- ♣ One hundred replicates of Bernoulli data, of size $n = 100$, were generated from $x_i = (i - 0.5)/n$, $i = 1, \dots, n$, with the logit function $\eta(x) = 3\{10^5 x^{11}(1-x)^6 + 10^3 x^3(1-x)^{10}\} - 2$.

Cubic spline logistic regression estimates were computed, for λ on a grid $\log_{10} \lambda = -6(0.1)0$, and for λ selected by indirect and direct cross-validations. The symmetrized Kullback-Leibler loss $L(\eta, \eta_\lambda)$ were evaluated for all estimates.

Slide 6

Slide 7



Slide 8

Approximate Bayesian Confidence Intervals

Based on the penalized weighted least squares problem at the converged η_λ , approximate posterior means and standard deviations can be computed, and in turn approximate Bayesian confidence intervals can be constructed.

The prior on η is Gaussian, with a diffuse component in $\mathcal{N}_J = \mathcal{H}_0$ and proper components in \mathcal{H}_β , $\beta = 1, \dots, p$. The sampling distributions of $\mathbf{Y}|\eta$ are generally non-quadratic in η .

Under certain parameter designation, the fitted values at the sampling points are actually the posterior mode.

The normal approximation of the posterior distribution is simply an application of the so-called Laplace approximation of integrals.

Software: gssanova in gss

With syntax similar to `glm`, `gssanova` can be used to fit non-Gaussian regression models using the penalized likelihood method. The families implemented are *binomial*, *Poisson*, *gamma*, *inverse Gaussian*, and *negative binomial*.

Slide 9

Only one link is implemented for each family: *logit* for binomial and negative binomial, and *log* for Poisson, gamma, and inverse Gaussian. The binomial and Poisson links are canonical.

The dispersion $a(\phi)$ is known for binomial, Poisson, and negative binomial, unknown for gamma and inverse Gaussian.

Poisson regression may be used to fit distributional models.

Gamma regression may be used to estimate spectral densities of stationary time series.