Exponential Families and GLMs
(27.02 current version)

ginsbourger@stat.unibe.ch

Spring 2017

About Maximum Likelihood Estimation

Major questions that arise regarding MLE notably include

- for a given model, the existence of some likelihood maximizer(s),
- and, when existing, how to compute the MLE in practice.

Moreover, questions such as the consistency and further asymptotic properties of MLE are of utmost importance both in theory and in practice.

When the likelihood function is sufficiently regular (say twice continuously differentiable over an open $\Theta \subset \mathbb{R}^p$), the MLE is typically searched among the roots of the so-called *score function* (here $\boldsymbol{y}$ is treated as fixed):

$$U : \boldsymbol{\theta} \in \Theta \to U(\boldsymbol{\theta}) = \nabla \ell(\boldsymbol{\theta}; \boldsymbol{y}) = \left(\frac{\delta}{\delta \theta_j} \ell(\boldsymbol{\theta}; \boldsymbol{y}) \right)_{i=1,\ldots,p}.$$

Now, finding solutions to $U(\boldsymbol{\theta}) = \mathbf{0}$ (called "normal equation") is in most cases non-trivial as U is often non-linear. Iterative procedures are then used to approach (potential) solutions.

About Maximum Likelihood Estimation

Another function that comes into play in maximum likelihood estimation is

$$\mathcal{J} : \boldsymbol{\theta} \in \Theta \to \mathcal{J}(\boldsymbol{\theta}) = -\nabla^2 \ell(\boldsymbol{\theta}; \boldsymbol{y}),$$

the so-called "observed information matrix".

About Maximum Likelihood Estimation

$\mathcal{J}$ can notably be used to check if a solution to the normal equations actually is a local maximizer of ℓ. Also, $\mathcal{J}$ is instrumental in numerical MLE procedures (Cf. Newton-Raphson algorithms). The "expected version" of $\mathcal{J}$ further plays a key role in asymptotic results regarding MLE's variance.

Note that the existence of solutions, the convergence of algorithms, asymptotic properties of MLE are model-dependent and require serious examination. We will now present a distribution class that will be very useful in logistic and Poisson regression and for which things are (relatively) simple.

Preliminaries

Let $n \in \mathbb{N}\backslash\{0\}$ and ν be a σ-finite measure on $\mathbb{R}^n$ (equipped with its Borel σ-field). Then one defines the *natural parameter space* as follows:

$$D_\nu = \left\{ \boldsymbol{\alpha} \in \mathbb{R}^n : \int_{\mathbb{R}^n} \exp(\langle \boldsymbol{\alpha}, \boldsymbol{y} \rangle) \mathrm{d}\nu(\boldsymbol{y}) < \infty \right\}.$$

The function $\kappa_\nu : A \to \mathbb{R}$ defined by

$$\kappa_\nu(\boldsymbol{\alpha}) = \log \left(\int_{\mathbb{R}^n} \exp(\langle \boldsymbol{\alpha}, \boldsymbol{y} \rangle) \mathrm{d}\nu(\boldsymbol{y}) \right)$$

(with $\log(\infty) = \infty$) is called the *cumulant (generating) function*.

Two convexity properties

The set D_ν is convex. Furthermore the function κ_ν is convex.

Definition

Natural Exponential Families (NEFs)

A family $(P_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha} \in A \subset \mathbb{R}^n}$ of probability distributions on $\mathbb{R}^n$ possessing densities $p_{\boldsymbol{\alpha}}(\boldsymbol{y})$ with respect to ν defines a NEF if

$$A \subset D_\nu \text{ and}$$
$$f_{\boldsymbol{\alpha}}(\boldsymbol{y}) = \exp\left(\langle \boldsymbol{\alpha}, \boldsymbol{y} \rangle - \kappa_\nu(\boldsymbol{\alpha})\right) \text{ for } \boldsymbol{\alpha} \in A and \boldsymbol{y} \in \mathbb{R}^n.$$

The parameters $\boldsymbol{\alpha} \in A$ are referred to as *canonical parameters*.

A NEF is said *saturated* (or *full*) when $A = D_\nu$ and *regular* when D_ν is non-empty and open.

A NEF is said *minimal* when the support of ν is not included in an $n-1$-dimensional subspace of $\mathbb{R}^n$. Note that if a NEF is minimal, then the cumulant function is known to be strictly convex.

Note on a common generalization

Note that the terminology of exponential families suffers from heterogeneity across the literature. Here natural refers to the fact that $\boldsymbol{y}$ is not transformed. In more general settings, exponential families are often defined with k-dimensional α and $\boldsymbol{y}$ is replaced by $T(\boldsymbol{y})$ in the exponential, where $T : \mathbb{R}^n \to \mathbb{R}^k$ and $T(\boldsymbol{y})$ is referred to as *sufficient statistic*. Here and in the following we will assume by default that $k = n$ and T is the identity of $\mathbb{R}^n$.

Some properties of regular NEFs

For regular NEFs, the cumulant function κ_ν is infinitely differentiable on D_ν and one can show that the moments of $P_{\boldsymbol{\alpha}}$ can be expressed in terms of the derivatives of κ_ν. In particular, the following holds:

Proposition

Let $\boldsymbol{Y}$ be a random vector which distribution belongs to a regular NEF $\{P_{\boldsymbol{\alpha}}; \boldsymbol{\alpha} \in A\}$ and $\boldsymbol{\alpha} \in A$. Then its mean $\boldsymbol{\mu}$ and covariance matrix $\Sigma_{\boldsymbol{\alpha}}$ satisfy

$$\mathbb{E}_{\boldsymbol{\alpha}}[Y] = \nabla \kappa_\nu(\boldsymbol{\alpha}) \text{ and}$$
$$\Sigma_{\boldsymbol{\alpha}} = \mathrm{Cov}_{\boldsymbol{\alpha}}[Y] = \nabla^2 \kappa_\nu(\boldsymbol{\alpha})$$

For regular minimal NEFs, the map $\boldsymbol{m} : \boldsymbol{\alpha} \in A \to \boldsymbol{m}(\boldsymbol{\alpha}) = \mathbb{E}_{\boldsymbol{\alpha}}[Y] \in \mathbb{R}^n$ is actually a diffeomorphism between A and $\mathcal{M} = \{\mathbb{E}_{\boldsymbol{\alpha}}[Y]; \boldsymbol{\alpha} \in A\}$. This leads to an alternative parametrization of the NEF (via $\boldsymbol{\mu}$ instead of $\boldsymbol{\alpha}$).

Examples of NEFs: binomial distributions

Let us first consider $B(m, \rho)$ with $m \geq 1$ and $\rho \in (0, 1)$. This distribution on $\{0, 1, \ldots, m\}$ is dominated by $\nu = \sum_{k=0}^{m} \binom{m}{k} \delta_{\{k\}}$.

The density (in the sense of a Radon-Nikodym derivative) of $B(m, \rho)$ with respect to ν is given by

$$p(y) = \exp\left(y \log\left(\frac{\rho}{1-\rho}\right) - m \log\left(\frac{1}{1-\rho}\right)\right).$$

Hence by posing $\alpha = \log\left(\frac{\rho}{1-\rho}\right)$ and $\kappa_\nu(\alpha) = m \log(1 + \exp(\alpha))$, binomial distributions appear as a regular NEF. Furthermore, differentiating the cumulant function one recovers indeed

$$\mathbb{E}[Y] = \frac{\mathrm{d}}{\mathrm{d}\alpha}\kappa_\nu(\alpha) = \frac{m \exp(\alpha)}{1 + \exp(\alpha)} = m\rho$$

Example of NEFs: Tensorization

The previously considered example (Binomial) is scalar (n=1).

Considering now a sequence $Y_i, i = 1, \ldots, n$ of independent random variables with binomial distributions $B(m, \rho_i)$ ($\rho_i \in (0, 1)$), the distribution of $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ is dominated by $\nu^{\otimes n}$. It admits the following density:

$$p(\boldsymbol{y}; \boldsymbol{\alpha}) = \exp(\langle \boldsymbol{\alpha}, \boldsymbol{y} \rangle - \boldsymbol{\kappa}_\nu(\boldsymbol{\alpha})),$$

where $\boldsymbol{\alpha} = \begin{pmatrix} \log\left(\frac{\rho_1}{1-\rho_1}\right) \\ \vdots \\ \log\left(\frac{\rho_n}{1-\rho_n}\right) \end{pmatrix}$ and $\boldsymbol{\kappa}_\nu(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \kappa_\nu(\alpha_i)$.

Examples of NEFs: Poisson distributions

For $\lambda > 0$, the Poisson distribution is dominated by the discrete measure $\nu = \sum_{k \geq 0} \frac{1}{k!} \delta_k$. The density of $P(\lambda)$ with respect to ν is given by

$$f(y) = \exp(y \log(\lambda) - \lambda).$$

The family $\{P_\lambda; 0 < \lambda < \infty\}$ constitutes a regular NEF with

$$\alpha = \log(\lambda) \text{ and } \kappa_\nu(\alpha) = \exp(\alpha).$$

Examples of NEFs: Gaussian distributions

The family $\{N(\mu, \sigma^2) : \mu \in \mathbb{R}\}$ defines for known σ^2 a regular NEF by taking for ν the Gaussian measure $N(0, \sigma^2)$, for canonical parameter $\alpha = \mu/\sigma^2$ and for cumulant function $\kappa_\nu(\alpha) = \frac{1}{2}\sigma^2\alpha^2$.

More generally, for Σ an invertible covariance matrix the family $\{N_n(\boldsymbol{\mu}, \Sigma) : \boldsymbol{\mu} \in \mathbb{R}^n\}$ defines a NEF characterized by

$$\boldsymbol{\alpha} = \Sigma^{-1}\boldsymbol{\mu} \text{ and}$$

$$\kappa_\nu(\boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\alpha}'\Sigma\boldsymbol{\alpha}$$

with $N_n(\mathbf{0}, \Sigma)$ as dominating measure.

Maximum Likelihood Estimation for NEFs

In the case of NEFs, the log-likelihood writes

$$\ell(\boldsymbol{\alpha};\boldsymbol{y}) = \langle \boldsymbol{\alpha}, \boldsymbol{y}\rangle - \kappa_\nu(\boldsymbol{\alpha})$$

As κ_ν is convex, ℓ is then a concave function. Under minimality of the NEF, strict concavity ensures unicity of the MLE in case it exists.

Furthermore, the normal equations associated with a NEF are

$$\nabla\kappa_\nu(\boldsymbol{\alpha}) = \mathbb{E}_{\boldsymbol{\alpha}}[\boldsymbol{Y}] = \boldsymbol{y}.$$

It can be guaranteed that this equation possesses a solution by assuming a regular NEF and that $\boldsymbol{y}$ belongs to the interior of the closure K of H, the convex hull of ν's support (Cf. for instance Antoniadis, Berruyer and Carmona's 1992 book).

A detour through curved EFs

A family $\{Q_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \Theta\}$ of probability distributions defines a curved EF it it can be smoothly embedded in a saturated regular NEF in the following sense:

> Let Θ be a nonempty open set in $\mathbb{R}^p$ and $\boldsymbol{\tau}$ be a C^2-diffeomorphism from Θ to the natural parameter space A of a n-dimensional saturated, minimal, regular NEF $\mathcal{P}$ dominated by ν. One calls curved Exponential Family of dimension p (and of order n) defined by $\boldsymbol{\tau}$ the sub-family $\{P_{\boldsymbol{\tau}(\boldsymbol{\theta})}; \boldsymbol{\theta} \in \Theta\}$. If the diffeomorphism $\boldsymbol{\tau}$ is linear, the EF is said *flat* or *linear*.

In particular, for a curved EF, for any $\boldsymbol{y} \in \mathbb{R}^n$ and $\boldsymbol{\theta} \in \Theta$ we have

$$L(\boldsymbol{\theta};\boldsymbol{y}) = \exp\{\langle \boldsymbol{\tau}(\boldsymbol{\theta}), \boldsymbol{y}\rangle - \kappa_\nu(\boldsymbol{\tau}(\boldsymbol{\theta}))\}$$

and

$$\ell(\boldsymbol{\theta};\boldsymbol{y}) = \langle \boldsymbol{\tau}(\boldsymbol{\theta}), \boldsymbol{y}\rangle - \kappa_\nu(\boldsymbol{\tau}(\boldsymbol{\theta}))$$

A detour through curved EFs: non-linear regression

Let us consider the non-linear regression framework for a first example. In the case where

$$\boldsymbol{Y} = \boldsymbol{\tau}(\theta) + \varepsilon,$$

where $\boldsymbol{\tau}$ (standing for the previously named $\boldsymbol{v}$ function) satisfies the hypotheses above and the error ε is assumed to have distribution $\nu = \mathcal{N}(\boldsymbol{0}, \sigma^2 I_n)$ (with $\sigma^2 > 0$), the random vector $\boldsymbol{Y}$ possesses the density

$$p_{\boldsymbol{\theta}}(\boldsymbol{y}) = \exp\left\{\frac{1}{\sigma^2}\langle \boldsymbol{\tau}(\boldsymbol{\theta}), \boldsymbol{y}\rangle - \frac{1}{2\sigma^2}||\boldsymbol{\tau}(\boldsymbol{\theta})||^2\right\},$$

with respect to ν. Hence this Gaussian non-linear regression model appears as a special case of a curved Exponential Family.

A detour through curved EFs

Coming back to general curved EFs, let us now focus on MLE. From

$$\ell(\boldsymbol{\theta};\boldsymbol{y}) = \langle \boldsymbol{\tau}(\boldsymbol{\theta}), \boldsymbol{y}\rangle - \kappa_\nu(\boldsymbol{\tau}(\boldsymbol{\theta})),$$

we obtain the score

$$U(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta};\boldsymbol{y}) = \langle \nabla_{\boldsymbol{\theta}}\boldsymbol{\tau}(\boldsymbol{\theta}), \boldsymbol{y}\rangle - \nabla_{\boldsymbol{\theta}}\boldsymbol{\tau}(\boldsymbol{\theta})^T \nabla_{\boldsymbol{\tau}}\kappa_\nu(\boldsymbol{\tau}(\boldsymbol{\theta})).$$

Hence under our regularity assumptions the MLE $\widehat{\boldsymbol{\theta}}$ is solution to

$$\nabla_{\boldsymbol{\theta}}\boldsymbol{\tau}(\widehat{\boldsymbol{\theta}})^T\left(\mathbf{y} - \nabla_{\boldsymbol{\tau}}\kappa_\nu(\boldsymbol{\tau}(\widehat{\boldsymbol{\theta}}))\right) = \mathbf{0}.$$

Using that $\boldsymbol{\mu}(\boldsymbol{\theta}) \equiv \boldsymbol{\mu}(\boldsymbol{\tau}(\boldsymbol{\theta})) = \mathbb{E}_{\boldsymbol{\tau}(\boldsymbol{\theta})}(\boldsymbol{Y}) = \nabla_{\boldsymbol{\tau}}\kappa_\nu(\boldsymbol{\tau}(\widehat{\boldsymbol{\theta}}))$ and $\nabla_{\boldsymbol{\theta}}\boldsymbol{\mu}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{\tau}(\boldsymbol{\theta})}(\boldsymbol{Y}) = \Sigma_{\boldsymbol{\tau}(\boldsymbol{\theta})}\nabla_{\boldsymbol{\theta}}\boldsymbol{\tau}(\boldsymbol{\theta})$, whereof

$$U(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\boldsymbol{\mu}(\boldsymbol{\theta})^T\Sigma^{-1}_{\boldsymbol{\tau}(\boldsymbol{\theta})}\left(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\theta})\right),$$

the latter normal equation can be reformulated as

$$\nabla_{\boldsymbol{\theta}}\boldsymbol{\mu}(\widehat{\boldsymbol{\theta}})^T\Sigma^{-1}_{\boldsymbol{\tau}(\widehat{\boldsymbol{\theta}})}\left(\mathbf{y} - \boldsymbol{\mu}(\widehat{\boldsymbol{\theta}})\right) = \mathbf{0}.$$

Principles

In classical linear modelling it was assumed that the response Y has expectation $\mathbf{d}(X)^T\boldsymbol{\theta}$, where $\mathbf{d}(x) = (f_j(x;\boldsymbol{\theta}))_{j=1\dots p}$ for some basis functions $f_j : \mathcal{X} \to \mathbb{R}$.

In so-called *Generalized Linear Models* (GLMs), one rather assumes that $Y_i = h(\mathbf{d}(X)^T\boldsymbol{\theta})$ for some (regular) real-valued function h.

For instance, in the case of the logistic regression model, where $\mathbb{E}[Y_i] = \rho(\boldsymbol{X}_i)$, the function h was defined by $h(x) = \exp(x)/(1+\exp(x))$. Let us now introduce GLMs in a more general settings (in the sense of the seminal 1972 article by *Nelder and Wedderburn* as quoted in ABC1992).

Definition of standard GLMs

A standard Generalized Linear Model is defined by

- a n-dimensional random vector $\boldsymbol{Y}$ having independent components with distribution from a saturated regular Natural Exponential Family,
- a full column-rank design matrix $D \in \mathbb{R}^{n \times p}$,
- a p-dimensional vector $\boldsymbol{\theta}$ belonging a non-empty open subset $\Theta \subset \mathbb{R}^p$,
- a diffeomorphism $h : \mathbb{R} \mapsto \mathbb{R}$ such that

$$\mathbb{E}[Y_i] = h(\mathbf{d}_i^T\boldsymbol{\theta}) = h((D\boldsymbol{\theta})_i) \quad 1 \leq i \leq n,$$

  where $\mathbf{d}_i^T$ is the ith line of D.

The function $g = h^{-1}$ is called *link function*.

Remarks on the definition of standard GLMs

Note that from the assumptions that $\boldsymbol{Y}$ stems from a NEF and has mean $\mathbf{h}(D\boldsymbol{\theta})$ with $\mathbf{h} : \mathbb{R}^n \mapsto \mathbb{R}^n$ defined by $(\mathbf{h}(\boldsymbol{x}))_i = h(x_i)$ for $1 \leq i \leq n$, the parametrization by the mean allows to represent $\boldsymbol{Y}$'s density in the form:

$$p(\boldsymbol{y};\boldsymbol{\theta}) = \exp\left\{\boldsymbol{m}^{-1}(\mathbf{h}(D\boldsymbol{\theta}))^T\boldsymbol{y} - \kappa_\nu(\boldsymbol{m}^{-1}(\mathbf{h}(D\boldsymbol{\theta})))\right\},$$

which possesses in turn a curved EF structure with

$$\boldsymbol{\tau}(\boldsymbol{\theta}) = \boldsymbol{m}^{-1}(\mathbf{h}(D\boldsymbol{\theta})).$$

Often the link function is chosen in accordance with the NEF underlying the standard GLM. Given the NEF and its $\boldsymbol{m}$ function, one commonly uses the so-called *canonical link function* defined by the equation ($\mathbf{g}$ as $\mathbf{h}$ above)

$$\mathbf{g}(\boldsymbol{\mu}) = \boldsymbol{\alpha} = \boldsymbol{m}^{-1}(\boldsymbol{\mu}),$$

i.e., in other words, $\mathbf{h} = \boldsymbol{m}$ (here same 1d NEF assumed for all components). With this link function, $\boldsymbol{\tau}$ is linear and the overall NEF is hence flat.

GLMs with dispersion parameter

Beyond the standard GLM, densities f_{α_i} used in practice for the marginal distributions sometimes depend on a parameter ϕ, called *dispersion parameter*, and one obtains the more general definition:

A GLM with dispersion parameter is a GLM in which the condition specifying $\boldsymbol{Y}$'s distribution is replaced by: the density of Y_i is of the form

$$f_{\alpha_i}(y_i) = \exp\{(\alpha_i y_i - b(\alpha_i)/a(\phi) + c(y_i,\phi)\},$$

where a, b, c are given real functions.

Likelihood of a GLM with dispersion parameter

The log-likelihood function of $\boldsymbol{\alpha}$ (given a realization $\boldsymbol{y}$ of a random $\boldsymbol{Y}$ vector from a GLM with dispersion parameter ϕ) is the given by

$$
\begin{aligned}
L(\boldsymbol{\alpha};\boldsymbol{y}) &= \prod_{i=1}^{n} \exp\{(\alpha_i y_i - b(\alpha_i)/a(\phi) + c(y_i,\phi)\} \\
&= \exp\left(\frac{\langle\boldsymbol{\alpha},\boldsymbol{y}\rangle - \bar{b}(\boldsymbol{\alpha})}{a(\phi)} + \bar{c}(\boldsymbol{y},\phi)\right),
\end{aligned}
$$

where $\bar{b}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} b(\alpha_i)$ and $\bar{c}(\boldsymbol{y},\phi) = \sum_{i=1}^{n} c(y_i,\phi)$.

Hence one obtains for the log-likelihood function:

$$
\ell(\boldsymbol{\alpha};\boldsymbol{y}) = \frac{\langle\boldsymbol{\alpha},\boldsymbol{y}\rangle - \bar{b}(\boldsymbol{\alpha})}{a(\phi)} + \bar{c}(\boldsymbol{y},\phi).
$$

About the “dispersion” parameter

For fixed ϕ, an affine change of variable highlights that the underlying distribution actually stems from an exponential family.

This leads to:

$$
\begin{aligned}
\mathbb{E}[\boldsymbol{Y}] &= \nabla\bar{b}(\boldsymbol{\alpha}), \\
\mathrm{Var}[\boldsymbol{Y}] &= a(\phi)\nabla^2\bar{b}(\boldsymbol{\alpha}).
\end{aligned}
$$

The last equation illustrates why ϕ is called the dispersion parameter.

Examples of GLMs: logistic regression model

As seen previously in the logistic regression model, the expectation of the response Y_i (i.e. the probability of its value being 1) is related to the covariables through the link function $g(\rho) = \log(\rho/(1-\rho))$ with

$$
\rho(\boldsymbol{x}_i) = (\exp(\mathbf{d}(\boldsymbol{x}_i)^T\boldsymbol{\theta}))/(1+\exp(\mathbf{d}(\boldsymbol{x}_i)^T\boldsymbol{\theta}))
$$

Remembering that for the Binomial model $\boldsymbol{\alpha} = \left(\log\left(\frac{\rho_i}{1-\rho_i}\right)\right)$ and $\bar{\kappa}_\nu(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \kappa(\alpha_i)$ where $\kappa(\alpha) = m\log(1+\exp(\alpha))$, here

$$
\ell(\boldsymbol{\theta};\boldsymbol{y}) = \langle\boldsymbol{\alpha},\boldsymbol{y}\rangle - \bar{\kappa}_\nu(\boldsymbol{\alpha}) = \sum_{i=1}^{n}(D\boldsymbol{\theta})_i y_i - \sum_{i=1}\log(1+\exp((D\boldsymbol{\theta})_i).
$$

Note that other link functions are also sometimes used, such as the

- *probit* function $g(\rho) = \Phi^{-1}(\rho)$ (where Φ is the c.d.f. of $\mathcal{N}(0,1)$),
- complementary log-log function $g(\rho) = \log(-\log(1-\rho))$,
- *power* (“Box-Cox”) function $g(\rho) = \frac{\rho^\gamma - 1}{\gamma}$ for $\gamma > 0$ and $\log(\rho)$ for $\gamma = 0$.

Examples of GLMs: Poisson regression model

For the Poisson regression model, the response is assumed to follow a Poisson distribution with intensity parameter $\lambda(\boldsymbol{x}) = \exp\left(\mathbf{d}(\boldsymbol{x})^T\boldsymbol{\theta}\right)$.

Hence here the link function is $g(\lambda) = \log(\lambda)$ and the log-likelihood function is given by

$$
\ell(\boldsymbol{\theta};\boldsymbol{y}) = \sum_{i=1}^{n}(D\boldsymbol{\theta})_i y_i - \sum_{i=1}^{n}\exp\left((D\boldsymbol{\theta})_i\right).
$$

Examples of GLMs: Gaussian linear regression model

When the error variance σ^2 is known, the Gaussian linear regression model is a particularly simple GLM with identity link function.

When σ^2 is unknown, one can expand the density of $\mathcal{N}(\mu, \sigma^2)$ as follows:

$$\begin{aligned} p(y; \mu, \sigma^2) &= \exp\left\{ \left(y\mu - \frac{\mu^2}{2} \right) / \sigma^2 + \left(-\frac{y^2}{2\sigma^2} - \frac{\log(2\pi\sigma^2)}{2} \right) \right\} \\ &= \exp\{(\alpha y - b(\alpha)/a(\phi) + c(y_i, \phi)\}, \end{aligned}$$

with $\alpha = \mu$, $b(\alpha) = \mu^2/2$, $a(\phi) = \sigma^2$, and $c(y, \phi) = -\frac{y^2}{2\sigma^2} - \frac{\log(2\pi\sigma^2)}{2}$.

Taking the identity as link function, one defines a GLM with dispersion parameter. In this model, the log-likelihood function is

$$\ell(\mu, \sigma^2; \mathbf{y}) = \frac{1}{\sigma^2} \left\{ \sum_{i=1}^{n} (D\boldsymbol{\theta})_i y_i - \sum_{i=1}^{n} \frac{(D\boldsymbol{\theta})_i}{2} \right\} - \sum_{i=1}^{n} \frac{y_i}{2\sigma^2} - m\frac{\log(2\pi\sigma^2)}{2}.$$

Estimation of GLM regression coefficients (1)

Let us now consider a GLM regression setting and assume that, if any, the dispersion parameter ϕ is fixed, so that the GLM has a curved EF of dimension p associated with it. Indeed recall that

$$p(\boldsymbol{y}; \boldsymbol{\theta}) = \exp\left\{ \boldsymbol{m}^{-1}(\mathbf{h}(D\boldsymbol{\theta}))^T \boldsymbol{y} - \kappa_\nu(\boldsymbol{m}^{-1}(\mathbf{h}(D\boldsymbol{\theta}))) \right\},$$

with $\mathbf{h} : \mathbb{R}^n \mapsto \mathbb{R}^n$ defined by $(\mathbf{h}(\boldsymbol{x}))_i = h(x_i)$ for $1 \leq i \leq n$, so that with the former curved EF notation we have

$$\boldsymbol{\tau}(\boldsymbol{\theta}) = \mu^{-1}(\mathbf{h}(D\boldsymbol{\theta})).$$

From there, revisiting equations satisfied by MLE in curved EF gives

$$D^T \nabla_{\boldsymbol{\theta}} \mathbf{h}(D\widehat{\boldsymbol{\theta}})^T \Sigma^{-1}_{\boldsymbol{\tau}(\widehat{\boldsymbol{\theta}})} \left(\mathbf{y} - \boldsymbol{m}(\widehat{\boldsymbol{\theta}}) \right) = \mathbf{0}.$$

where we have used that $\nabla \mu(\boldsymbol{\theta}) = \nabla \mathbf{h}(\boldsymbol{\theta}) D$.

Estimation of GLM regression coefficients (2)

For the case where the link is canonical, we obtain with $\boldsymbol{\tau}(\boldsymbol{\theta}) = D\boldsymbol{\theta}$

$$D^T(\boldsymbol{y} - \boldsymbol{m}(\widehat{\boldsymbol{\theta}})) = \mathbf{0}.$$

For the case where the link function is the identity then $\boldsymbol{\tau}(\boldsymbol{\theta}) = \mu(D\boldsymbol{\theta})$ and the MLE equation becomes

$$D^T \Sigma^{-1}_{\boldsymbol{\tau}(\widehat{\boldsymbol{\theta}})} \left(\mathbf{y} - D\widehat{\boldsymbol{\theta}} \right) = \mathbf{0}.$$

In any case, iterative methods are appealed to in order to (approximately) solve for the MLE in such models.