GSI2015 2nd conference on Geometric Science of Information
28-30 Oct 2015 Ecole Polytechnique Paris-Saclay

Second-order Optimization over the Multivariate Gaussian Distribution

Luigi Malagò [1] [2] Giovanni Pistone

[1]Shinshu University JP & INRIA Saclay FR

[2]de Castro Statistics, Collegio Carlo Alberto, Moncalieri IT

Introduction

- This is is the presentation by Giovanni of the paper with the same title in the Proceedings.
- Unfortunately, Giovanni is the least qualified of the two authors to present this specific application of Information Geometry, his specific field of expertise being non-parametric Information Geometry and its applications in Probability and Statistical Physics. Luigi is currently working in Japan and could not make it.
- Among the two of us, Luigi is the responsible for the idea of using gradient methods and later, Newton methods, in black box optimization. Our collaboration started with the preparation of the FOGA 2011 paper

- L. Malagò, M. Matteucci, and G. Pistone. Towards the geometry of estimation of distribution algorithms based on the exponential family.

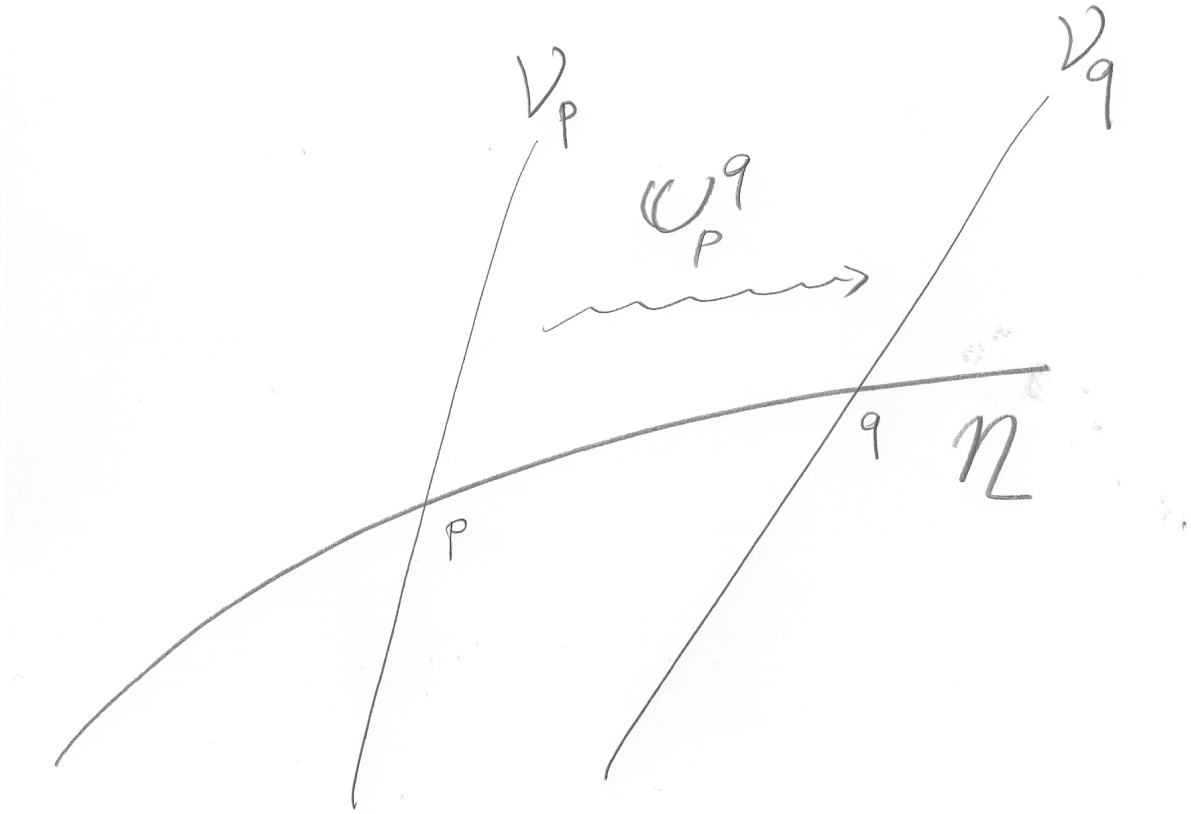
  In *Proceedings of the 11th workshop on Foundations of genetic algorithms*, FOGA '11, pages 230–242, New York, NY, USA, 2011. ACM

Summary

1. Geometry of the Exponential Family
2. Second-Order Optimization: The Newton Method
3. Applications to the Gaussian Distribution
4. Discussion and Future Work

- A short introduction for Taylor formulæ on Gaussian exponential families is provided. The binary case has been previously discussed in
- L. Malagò and G. Pistone. Combinatorial optimization with information geometry: Newton method. *Entropy*, 16:4260–4289, 2014.
- Riemannian Newton methods are discussed in a Session of this Conference cf,
- P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, Princeton, NJ, 2008.
  With a foreword by Paul Van Dooren
- The focus of this short presenation is on a specific framework for Information Geometry we call *statistical bundle*.

Hilbert vs Tangent vs Statistical Bundle



- S. Amari. Dual connections on the Hilbert bundles of statistical models.
  In *Geometrization of statistical theory (Lancaster, 1987)*, pages 123–151, Lancaster, 1987. ULDM Publ

- R. E. Kass and P. W. Vos. *Geometrical foundations of asymptotic inference.*
  Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, Inc., New York, 1997.

Statistical Bundle: Gaussian case

- $H_\alpha(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^m$, are Hermite polynomials of order 1 and 2.
- E.g, $m = 3$, $H_{010}(\boldsymbol{x}) = x_2$, $H_{011}(\boldsymbol{x}) = x_2 x_3$, $H_{020}(\boldsymbol{x}) = x_2^2 - 1$.
- The Gaussian model with sufficient statistics $\mathcal{B} = \{X_1, \ldots, X_n\} \subset \{H_\alpha | |\alpha| = 1, 2\}$, is

$$\mathcal{N} = \left\{ p(\boldsymbol{x}; \boldsymbol{\theta}) = \exp\left(\sum_{j=1}^{n} \theta_j X_j - \psi(\boldsymbol{\theta}) \right) \right\}$$

- The *fibers* are $\mathcal{V}_p = \text{Span}\left(X_j - \mathbb{E}_p\left[X_j\right] | j = 1, \ldots, n\right)$
- The *statistical bundle* is

$$S\mathcal{N} = \{(p, U) | p \in \mathcal{N}, U \in \mathcal{V}_p\}$$

- Each $U \in \mathcal{V}_p$, $p \in \mathcal{N}$, is a polynomial of degree up to 2 and $t \mapsto \mathbb{E}_q\left[\mathrm{e}^{tU}\right]$ is finite around 0, $q \in \mathcal{N}$
- Every polynomial X belongs to $\cap_{q \in \mathcal{N}} L^2(q)$

Parallel transports

Definition

- *e-transport*:
$$ {}^{e}\mathbb{U}_{p}^{q} \colon \mathcal{V}_p \ni U \mapsto U - \mathbb{E}_q\left[U\right] \in \mathcal{V}_q \ . $$
- *m-transport*: for each $U \in \mathcal{V}_p$ and $V \in \mathcal{V}_q$
$$ \left\langle U, {}^{m}\mathbb{U}_{q}^{p} V \right\rangle_p = \left\langle {}^{e}\mathbb{U}_{p}^{q} U, V \right\rangle_q $$

Properties

- ${}^{e}\mathbb{U}_{q}^{r}\, {}^{e}\mathbb{U}_{p}^{q} = {}^{e}\mathbb{U}_{p}^{r}$
- ${}^{m}\mathbb{U}_{q}^{r}\, {}^{m}\mathbb{U}_{p}^{q} = {}^{m}\mathbb{U}_{p}^{r}$
- $\left\langle {}^{e}\mathbb{U}_{p}^{q} U, {}^{m}\mathbb{U}_{p}^{q} V \right\rangle_q = \left\langle U, V \right\rangle_p$
- If $\frac{q}{p} V \in L^2(p)$, then ${}^{m}\mathbb{U}_{q}^{p} V$ is its orthogonal projection onto $\mathcal{V}_p$.

Parallel transports in coordinates I

We define on the statistical bundle $S\mathcal{N}$ a system of *moving frames*.

1. The *exponential frame* of the fiber $S_p\mathcal{N} = \mathcal{V}_p$ is the vector basis

$$\mathcal{B}p = \{X_j - \mathbb{E}_p[X_j] | j = 1, \ldots, n\}$$

2. Each element $U \in \mathcal{V}_p$ is uniquely written as

$$U = \sum_{j=1}^{n} \alpha_j(U)(X_j - \mathbb{E}_p[X_j]) = \boldsymbol{\alpha}(U)^T(\boldsymbol{X} - \mathbb{E}_p[\boldsymbol{X}])$$

3. The expression in the exponential frame of the scalar product is the Fisher information matrix:

$${}^e I_{ij}(p) = \langle X_i - \mathbb{E}_p[X_i], X_j - \mathbb{E}_p[X_j] \rangle_p = \mathrm{Cov}_p(X_i, X_j) = \frac{\partial^2}{\partial {\theta_i}^2} \theta_j \psi(\boldsymbol{\theta})$$

4.

$$U \mapsto \boldsymbol{\alpha}(U) = {}^e I(p)^{-1} \mathrm{Cov}_p(\boldsymbol{X}, U)$$

Parallel transports in coordinates II

5. The *mixture frame* of the fiber $S_p\mathcal{N} = \mathcal{V}_p$ is

$$ {}^e I(p)^{-1} \mathcal{B}p = \left\{ \sum_{i=1}^{n} {}^e I^{ij}(p)(X_i - \mathbb{E}_p[X_i]) \middle| j = 1, \ldots, n \right\} $$

6. Each element $V \in \mathcal{V}_p$ is uniquely written as

$$ V = \sum_{j=1}^{n} \beta_j(V) \sum_{i=1}^{n} {}^e I^{ij}(p)(X_i - \mathbb{E}_p[X_i]) = \boldsymbol{\beta}(V)^T \, {}^e I(p)^{-1}(\boldsymbol{X} - \mathbb{E}_p[\boldsymbol{X}]) $$

7. The coordinates in the mixture basis are given in matrix form by

$$ V \mapsto \boldsymbol{\beta}(V) = \mathrm{Cov}_p(\boldsymbol{X}, V) \ . $$

8. The matrix ${}^m I(p) = {}^e I(p)^{-1}$ is the matrix expression of the metric in the mixture frame.

$$ \boldsymbol{\alpha}(U) = {}^m I(p) \boldsymbol{\beta}(U), \quad \boldsymbol{\beta}(U) = {}^e I(p) \boldsymbol{\alpha}(U) \ . $$

Parallel transports in the moving frames

- The e-transport acts on the exponential coordinates as the identity,

$$\boldsymbol{\alpha}\left({}^{e}\mathbb{U}_{p}^{q}U\right)=\boldsymbol{\alpha}(U)$$

- Equivalently,

$$={}^{e}I\left(q\right)^{-1}\mathrm{Cov}_{q}\left(\boldsymbol{X},U\right)={}^{e}I\left(p\right)^{-1}\mathrm{Cov}_{p}\left(\boldsymbol{X},U\right)$$

- The m-transport acts on the mixture coordinates as the identity,

$$\boldsymbol{\beta}\left({}^{m}\mathbb{U}_{p}^{q}V\right)=\boldsymbol{\beta}(V)$$

REMARK

A *section* or *vector field* of the statistical bundle is a mapping $F:\mathcal{N}\ni p\mapsto F(p)\in\mathcal{V}_{p}$. As there are *two* distingushed charts on the model (exponential $p\mapsto\boldsymbol{\theta}(p)$ and mixture $p\mapsto\boldsymbol{\eta}(p)=\nabla\psi(\boldsymbol{\theta}(p))$) and *two* distinguished frames on each fiber, there are in general *four* distinguished expression of each section.

Score and statistical gradient

Definition

$t \mapsto p(t)$ is a curve in the model $\mathcal{N}$ and $f : \mathcal{N} \to \mathbb{R}$.

- The *score* of the curve $t \mapsto p(t)$ is a curve in the statistical bundle $t \mapsto (p(t), Dp(t)) \in S\mathcal{N}$ such that for all $X \in \mathrm{Span}\,(1, X_1, \ldots, X_n)$ it holds

$$\frac{d}{dt}\mathbb{E}_{p(t)}\left[X\right] = \left\langle X - \mathbb{E}_{p(t)}\left[X\right], Dp(t)\right\rangle_{p(t)}$$

- Usually,

$$Dp(t) = \frac{\dot{p}(t)}{p(t)} = \frac{d}{dt}\log p(t)$$

- The *statistical gradient* of $f : \mathcal{N} \to \mathbb{R}$ is a section of the statistical bundle, $p \mapsto (p, \mathrm{grad}\, f(p)) \in S\mathcal{N}$ such that for each regular curve $t \mapsto p(t)$, it holds

$$\frac{d}{dt} f(p(t)) = \langle \mathrm{grad}\, f(p(t)), Dp(t)\rangle_{p(t)}$$

Score and statistical gradient in coordinates

- Let the regular curve $t \mapsto p(t)$ be expressed in the exponential coordinates by $t \mapsto \boldsymbol{\theta}(t)$. The score $t \mapsto Dp(t)$ is expressed in the exponential frame by $t \mapsto \dot{\boldsymbol{\theta}}(t)$ that is,

$$Dp(t) = \sum_{j=1}^{n} \dot{\theta}_j(t)(X_j - \frac{\partial}{\partial \theta_j}\psi(\boldsymbol{\theta}(t)))$$

- Let the regular curve $t \mapsto p(t)$ be expressed in the mixture coordinates by $t \mapsto \boldsymbol{\eta}(t) = \nabla\psi(\boldsymbol{\theta}(t))$. The score is expressed in the mixture frame as $t \mapsto \dot{\boldsymbol{\eta}}(t)$.

- Let X be a random variable which belongs to all $L^2(p)$, $p \in \mathcal{N}$ and $f(p) = \mathbb{E}_p[f]$. Then $p \mapsto \operatorname{grad} f(p)$ exists and equals the orthogonal projection of X onto $\mathcal{V}_p$, namely

$$\operatorname{grad}(p \mapsto \mathbb{E}_p[X]) = {}^{e}I(p)^{-1} \operatorname{Cov}_p(\boldsymbol{X}, X)(\boldsymbol{X} - \mathbb{E}_p[\boldsymbol{X}]), \quad \boldsymbol{X} = (X_1, \ldots, X_n) .$$

- The expressions of $\operatorname{grad} f$ are of interest in optimization.

Taylor formula in the Statistical Bundle

- For a curve $t \mapsto p(t) \in \mathcal{N}$ connecting $p = p(0)$ to $q = p(1)$ and a function $f : \mathcal{N} \to \mathbb{R}$ the Taylor formula is

$$f(q) = f(p) + \frac{d}{dt} f(p(t))\Big|_{t=o} + \frac{1}{2} \frac{d^2}{dt^2} f(p(t))\Big|_{t=o} + R_2(f, p, q)$$

- The first derivative is computed with the statistical gradient and the score

$$f(q) = f(p) + \langle \operatorname{grad} f(p(0)), Dp(0) \rangle_p + \frac{1}{2} \frac{d}{dt} \langle \operatorname{grad} f(p(t)), Dp(t) \rangle_{p(t)}\Big|_{t=o} + R_2(f, p, q)$$

Accelleration and Hessian

$$\frac{d}{dt}\left\langle \operatorname{grad} f(p(t)), Dp(t)\right\rangle_{p(t)}\Big|_{t=o} =$$
$$\frac{d}{dt}\left\langle {}^{e}\mathbb{U}_{p(t)}^{p(0)} \operatorname{grad} f(p(t)), {}^{m}\mathbb{U}_{p(t)}^{p(0)} Dp(t)\right\rangle_{p(0)}\Big|_{t=o}$$

$$\frac{d}{dt}\left\langle \operatorname{grad} f(p(t)), Dp(t)\right\rangle_{p(t)}\Big|_{t=o} =$$
$$\frac{d}{dt}\left\langle {}^{m}\mathbb{U}_{p(t)}^{p(0)} \operatorname{grad} f(p(t)), {}^{m}\mathbb{U}_{p(t)}^{p(0)} Dp(t)\right\rangle_{p(0)}\Big|_{t=o}$$

$$\frac{d}{dt}\left\langle \operatorname{grad} f(p(t)), Dp(t)\right\rangle_{p(t)}\Big|_{t=o} =$$
$$\frac{d}{dt}\mathbb{E}_{p(0)}\left[\frac{p(t)}{p(0)} \operatorname{grad} f(p(t)) Dp(t)\right]\Big|_{t=o}$$

Accellerations

- Let us define the *acceleration* at t of a curve $t \mapsto p(t) \in \mathcal{N}$. The velocity is defined to be
  $t \mapsto (p(t), Dp(t)) = \left(p(t), \frac{d}{dt} \log\left(p(t)\right)\right) \in S\mathcal{N}$
- The *exponential acceleration* is

$$ {}^{e}\mathrm{D}^2 p(t) = \left. \frac{d}{ds} \, {}^{e}\mathbb{U}_{p(s)}^{p(t)} Dp(s) \right|_{s=t} $$

- The *mixture acceleration* is

$$ {}^{m}\mathrm{D}^2 p(t) = \left. \frac{d}{ds} \, {}^{m}\mathbb{U}_{p(s)}^{p(t)} Dp(s) \right|_{s=t} $$

- The *Riemannian acceleration* is

$$ {}^{0}\mathrm{D}^2 p(t) = \frac{1}{2} \left({}^{e}\mathrm{D}^2 p(t) + {}^{m}\mathrm{D}^2 p(t) \right) $$

Covariant derivatives I

- $p \mapsto (p, F(p))$, $p \mapsto (p, G(p))$, are sections of $S\mathcal{N}$, with expressions in the moving frames

$$F(p) = \sum_{j=1}^{n} \alpha^j(p)(X_j - \mathbb{E}_p[X_j]) \ ,$$

$$F(p) = \sum_{j=1}^{n} \beta^j(p) \sum_{i=1}^{n} {}^e I^{ij}(p)(X_i - \mathbb{E}_p[X_i]) \ ,$$

$$G(p) = \sum_{j=1}^{n} \gamma^j(p)(X_j - \mathbb{E}_p[X_j]) \ ,$$

$$G(p) = \sum_{j=1}^{n} \delta^j(p) \sum_{i=1}^{n} {}^e I^{ij}(p)(X_i - \mathbb{E}_p[X_i]) \ .$$

Covariant derivatives II

- The *exponential covariant derivative* is the vector field $p \mapsto (p, {}^{e}\mathrm{D}_{G}F(p))$, where

$$
\begin{aligned}
{}^{e}\mathrm{D}_{G}F(p) &= \sum_{j=1}^{n} \left\langle \operatorname{grad} \alpha^{j}(p), G(p) \right\rangle_{p} (X_j - \mathbb{E}_p [X_j]) \\
&= \sum_{j=1}^{n} \sum_{i=1}^{n} \gamma^{i}(p) (\partial_i \operatorname{grad} \alpha^{j}(p)) (X_j - \mathbb{E}_p [X_j])
\end{aligned}
$$

- The *mixture covariant derivative* is the vector field $p \mapsto (p, {}^{m}\mathrm{D}_{G}F(p))$, where

$$
{}^{m}\mathrm{D}_{G}F(p) = \sum_{j=1}^{n} \left\langle \operatorname{grad} \beta^{j}(p), G(p) \right\rangle_{p} \sum_{i=1}^{n} {}^{e}I^{ij}(p)(X_i - \mathbb{E}_p [X_i]) =
$$

$$
\sum_{j=1}^{n} \sum_{k=1}^{n} \gamma^{k}(p) \left\langle \operatorname{grad} \beta^{j}(p), X_k - \mathbb{E}_p [X_k] \right\rangle_{p} \sum_{i=1}^{n} {}^{e}I^{ij}(p)(X_i - \mathbb{E}_p [X_i])
$$

Covariant derivatives III

- The *Riemannian covariant derivative* is the vector field $p \mapsto (p, {}^0\mathrm{D}_G F(p))$ with

$$ {}^0\mathrm{D}_G F = \frac{1}{2} \left({}^e\mathrm{D}_G F + {}^m\mathrm{D}_G F \right) \ . $$

Hessians

- Let $f : \mathcal{N} \to \mathbb{R}$ be a mapping with gradient $p \mapsto (p, \operatorname{grad} f(p))$. Let $p \mapsto (p, G(p))$ be a vector field (section) of $S\mathcal{N}$.
- The *exponential Hessian* of f is the vector field $p \mapsto (p, {}^{e}\operatorname{Hess}_G f(p))$, with

$$ {}^{e}\operatorname{Hess}_G f(p) = {}^{e}\mathrm{D}_G \operatorname{grad} f(p) \ . $$

- The *mixture Hessian* of f is the vector field $p \mapsto (p, {}^{m}\operatorname{Hess}_G f(p))$, with

$$ {}^{m}\operatorname{Hess}_G f(p) = {}^{m}\mathrm{D}_G \operatorname{grad} f(p) \ . $$

- The *Riemannian Hessian* of F is the vector field $p \mapsto (p, {}^{0}\operatorname{Hess}_G F(p))$, with

$$ {}^{0}\operatorname{Hess}_G f(p) = {}^{0}\mathrm{D}_G \operatorname{grad} f(p) \ . $$

Taylor's formulæ I

1. $t \mapsto p(t)$ is the *mixture geodesic* connecting $p = p(0)$ to $q = p(1)$.

$$f(q) = f(p) + \langle \operatorname{grad} f(p), Dp(0) \rangle_p + \frac{1}{2} \left\langle {}^e\operatorname{Hess}_{Dp(0)} f(p), Dp(0) \right\rangle_p + R_2^+(p, q)$$

$$R_2^+(p, q) = \int_0^1 dt \left((1 - t) \left\langle {}^e\operatorname{Hess}_{Dp(t)} f(p(t)), Dp(t) \right\rangle_{p(t)} \right) - \frac{1}{2} \left\langle {}^e\operatorname{Hess}_{Dp(0)} f(p), Dp(0) \right\rangle_p$$

Taylor's formulæ II

2. $t \mapsto p(t)$ is the *exponential geodesic* connecting $p = p(0)$ to $q = p(1)$.

$$f(q) = f(p) + \langle \operatorname{grad} f(p), Dp(0) \rangle_p + \frac{1}{2} \left\langle {}^m\mathrm{Hess}_{Dp(0)} f(p), Dp(0) \right\rangle_p + R_2^-(p,q)$$

$$R_2^-(p,q) = \int_0^1 dt \left((1-t) \left\langle {}^m\mathrm{Hess}_{Dp(t)} f(p(t)), Dp(t) \right\rangle_{p(t)} \right) - \frac{1}{2} \left\langle {}^m\mathrm{Hess}_{Dp(0)} f(p), Dp(0) \right\rangle_p$$

Taylor's formulæ III

3. $t \mapsto p(t)$ is the *Riemannian geodesic* connecting $p = p(0)$ to $q = p(1)$.

$$f(q) = f(p) + \langle \operatorname{grad} f(p), Dp(0) \rangle_p + \frac{1}{2} \left\langle {}^0\mathrm{Hess}_{Dp(0)} f(p), Dp(0) \right\rangle_p + R_2^0(p, q)$$

where

$$R_2^0(p, q) = \int_0^1 dt(1-t) \left\langle {}^0\mathrm{Hess}_{Dp(t)} f(p(t)), Dp(t) \right\rangle_{p(t)} - \frac{1}{2} \left\langle {}^0\mathrm{Hess}_{Dp(0)} f(p), Dp(0) \right\rangle_p$$

Newton step

- Let $t \mapsto p(t)$ be the *exponential geodesic* starting at $p = p(0)$ with $Dp(0) = U$.
- Assume U is a critical point of

$$\mathcal{V}_{p(0)} \ni U \mapsto f(p) + \langle \operatorname{grad} f(p(0)), U \rangle_{p(0)} + \frac{1}{2} \langle {}^{m}\operatorname{Hess}_U f(p), U \rangle_{p(0)}$$

  that is

$$\operatorname{grad} f(p(0)) + {}^{m}\operatorname{Hess}_U f(p) = 0$$

- If $q = p(1)$, then

$$f(q) = f(p) - \frac{1}{2} \left\langle {}^{0}\operatorname{Hess}_U f(p), U \right\rangle_p + R_2^0(p, q)$$

Conclusion and work in progress

- Comparisons between the Riemannian Newton method e.g., Absil et al., and the statistical bundle setup are being performed.
- In particular, the use of alternative Hessians is of special interest.