Gaussianité des Erreurs et Méthodes de Maximum d'Entropie

Olivier Talagrand, Carlos Pires et Marc Bocquet

Troisième Colloque National sur l'Assimilation de Données
Grenoble
10 Décembre 2010

One can reasonably say that the ultimate purpose of assimilation is to achieve *bayesian estimation*, *i. e.* to determine the conditional probability distribution for the state of the atmospheric (or oceanic) flow, subject to the available information.

A large part of 'real life' assimilation algorithms (still ?) based on (heuristic extension to nonlinear situations of) statistical linear estimation, or *Best Linear Unbiased Estimator* (*BLUE*)

Data in the form

$$z = \Gamma x + \xi$$

Known data vector z belongs to *data space* $\mathcal{D}$, $dim\mathcal{D} = m$,
Unknown state vector belongs to *state space* S, $dimS = n$

Γ known (mxn)-matrix, ξ unknown 'error'

Best Linear Unbiased Estimator of x from z.

$$x^a = (\Gamma^T S^{-1} \Gamma)^{-1} \Gamma^T S^{-1} [z - \mu]$$
$$P^a \equiv E[(x^a - x)(x^a - x)^T] = (\Gamma^T S^{-1} \Gamma)^{-1}$$

where $\mu \equiv E(\xi)$
$S \equiv E\{[\xi - \mu][\xi - \mu]^T\}$

Requires *a priori* explicit knowledge of $E(\xi)$ and $E\{[\xi - E(\xi)][\xi - E(\xi)]^T\}$

Unambiguously defined iff rank$\Gamma = n$. *Determinacy condition*. Requires $m \geq n$.

If determinacy condition is verified, it is always possible to decompose data into

- A '*background*' estimate (*e. g.* forecast from the past), belonging to *state space*, with dimension n

$$x^b = x + \xi^b \; ; \quad E(\xi^b) = 0 \; ; \; E(\xi^b \xi^{b\mathrm{T}}) \equiv P^b$$

- An additional set of data (*e. g.* observations), belonging to *observation space*, with dimension $m - n = p$

$$y = Hx + \varepsilon \; ; \quad E(\varepsilon) = 0 \; ; \; E(\varepsilon \varepsilon^{\mathrm{T}}) \equiv R$$

with $E(\varepsilon \xi^{b\mathrm{T}}) = 0$

Then

$$x^a = x^b + P^b H^T [HP^b H^T + R]^{-1} (y - Hx^b)$$

$d \equiv y - Hx^b$ is *innovation vector*.

$$P^a = P^b - P^b H^T [HP^b H^T + R]^{-1} HP^b$$

Theoretical ‘justifications’

- In case ξ is gaussian, $\xi \sim \mathcal{N}[\mu, S]$, *BLUE* achieves bayesian estimation in the sense that $P(x \mid z) = \mathcal{N}[x^a, P^a]$

- Among all unbiased affine estimates of the form

$$x^a = \alpha + Az$$

BLUE minimizes quadratic estimation error $E[(x^a_i - x_i)^2]$ for any component x_i

- If only expectation and covariance matrix of ξ are known, gaussian pdf $p \sim \mathcal{N}[x^a, P^a]$ maximises entropy $- \int p \ln p \, dx = - E(\ln p(x))$, *i. e.* is ‘least committing’ estimate of x, while being compatible with μ and S.

Real reason

BLUE is simplest of non simplicist algorithms. Provides a systematic and consistent way for combining inhomogeneous data, taking into account their individual specified accuracies.

Only exceptions so far to *BLUE*

- *Ensemble Kalman Filter*, which does not need any linearity in dynamical evolution, but remains *BLUE* in 'updating' phase.

- *Particle filters* (P. J. van Leeuwen, R. Miller), which are bayesian without any need for linearity or gaussianity, but (still ?) do not exist in real meteorology or oceanography.

Within the world of data, innovation $d \equiv y - Hx^b = \varepsilon - H\zeta^b$ is only objective source of information on errors.

Question that can be objectively answered : is innovation gaussian ?

C. Pires, O. Talagrand and M. Bocquet, 2010, Diagnosis and impacts of non-Gaussianity of innovations in data assimilation, *Physica D*, **239**, 1701–1717.

Remark. Assuming data to be unbiased, together with $E(\varepsilon\zeta^{b\mathrm{T}}) = 0$, and setting $E(\zeta^b\zeta^{b\mathrm{T}}) \equiv P^b$, $E(\varepsilon\varepsilon^{\mathrm{T}}) \equiv R$ imply

$$E(d) = 0$$
$$E(dd^{\mathrm{T}}) = HP^bH^{\mathrm{T}} + R$$

which can be objectively checked against observed statistics of innovation (A. Weaver)

Study performed on innovations corresponding to satellite observations of brightness temperatures (instrument HIRS onboard NOAA 17), as available at ECMWF.

Statistics of the background errors, observation errors and innovations for the samples 4s, 5s, 6s, 7s, 11s, 12s, 14s 15s for 'ice-free' conditions and samples 5i, 6i, 7i, 14i and 15i for 'ice-covered' conditions. Table lists: (a) sample size N; (b) Specified standard deviation σ_{yo}, of the observation error, mean $m(\sigma_{yb})$ and standard deviation $\mathrm{std}(\sigma_{yb})$ of the specified standard deviation of the background error, innovation bias b_d, standard deviation σ_d of the innovation (all in K); (c) Skewness s_d, kurtosis k_d, and approximated negentropy J_d of the innovation with the 99%-statistically significant values of s_d, k_d and J_d marked in bold; (d) Heteroscedasticity measures h_{1b}, h_{1b} and h_{4b}, of σ_{yb}; (e) Maximum variance fraction β_g of the Gaussian error; (f) Bounds of the tuning factors: f_{o1}, f_{b1}, f_{o2} and f_{b2} of the specified standard deviation of errors. See text for details concerning quantities in (c), (d), (e) and (f).

Sample	4s	5s	6s	7s	11s	12s	14s	15s	5i	6i	7i	14i	15i
N	5907	4419	4132	3580	4426	6741	4081	4234	4980	4306	2043	3237	3581
σ_{yo}	0.6	0.6	0.6	0.75	1.0	2.0	0.5	0.6	0.6	0.6	0.75	0.5	0.6
$m(\sigma_{yb})$	0.083	0.126	0.243	0.432	0.904	1.040	0.354	0.207	0.378	1.031	2.502	1.442	0.618
$\mathrm{std}(\sigma_{yb})$	0.013	0.027	0.058	0.132	0.252	0.425	0.135	0.050	0.059	0.186	0.419	0.256	0.111
b_d	−0.01	−0.08	−0.13	−0.06	−0.02	0.17	−0.07	−0.04	0.03	−0.05	0.03	0.03	−0.01
σ_d	0.184	0.221	0.343	0.422	1.011	1.305	0.321	0.274	0.261	0.439	0.502	0.448	0.374
s_d	**−0.70**	**−0.63**	**−0.25**	0.10	**0.38**	**0.27**	−0.01	**−0.22**	**−0.34**	**−0.21**	**−0.26**	**−0.14**	**−0.21**
k_d	**1.02**	**0.90**	**0.35**	**0.56**	**0.65**	**2.07**	0.14	**0.23**	−0.07	**−0.47**	**−0.24**	**−0.33**	**−0.28**
J_d	**0.063**	**0.050**	**0.008**	**0.007**	**0.021**	**0.095**	0.0004	**0.005**	**0.010**	**0.008**	**0.007**	**0.004**	**0.005**
h_{1b}	0.99	0.98	0.97	0.96	0.96	0.93	0.93	0.97	0.99	0.98	0.99	0.98	0.98
h_{3b}	1.04	1.08	1.09	1.19	1.11	1.23	1.29	1.10	1.03	1.04	1.04	1.04	1.05
h_{4b}	1.10	1.23	1.28	1.61	1.33	1.72	2.08	1.32	1.10	1.11	1.10	1.12	1.12
β_g	0.646	0.675	0.852	0.983	0.810	0.965	1.000	0.843	0.637	0.566	0.622	0.643	0.638
f_{o1}	0.248	0.304	0.528	0.557	0.910	0.641	0.643	0.419	0.347	0.551	0.528	0.719	0.499
f_{b1}	1.762	1.409	1.266	0.925	0.969	1.141	0.848	1.182	0.544	0.316	0.156	0.245	0.477
f_{o2}	0.184	0.211	0.220	0.074	0.441	0.123	0.014	0.181	0.262	0.482	0.412	0.536	0.376
f_{b2}	1.305	0.978	0.527	0.122	0.470	0.219	0.018	0.511	0.411	0.276	0.122	0.183	0.359

Assumed independence of observation and background errors

$$(\sigma^d)^2 = (\sigma_{so})^2 + (\sigma_{sb})^2$$

In reality, rhs is much larger than the lhs. For most channels, $\sigma_{so} > \sigma^d$. At ECMWF, σ_{so} has been purposefully inflated in order to compensate for the neglect of spatial correlation of observation errors.

Gaussianity of innovation evaluated on the basis of

- skewness $s_d \equiv E(d'/\sigma^d)^3$
- kurtosis $k_d \equiv E(d'/\sigma^d)^4 - 3$

where E now denotes sample mean, and d' is unbiased innovation. Both skewness and kurtosis are 0 for a gaussian variable.

Statistics of the background errors, observation errors and innovations for the samples 4s, 5s, 6s, 7s, 11s, 12s, 14s 15s for 'ice-free' conditions and samples 5i, 6i, 7i, 14i and 15i for 'ice-covered' conditions. Table lists: (a) sample size N; (b) Specified standard deviation σ_{yo}, of the observation error, mean $m(\sigma_{yb})$ and standard deviation $\mathrm{std}(\sigma_{yb})$ of the specified standard deviation of the background error, innovation bias b_d, standard deviation σ_d of the innovation (all in K); (c) Skewness s_d, kurtosis k_d, and approximated negentropy J_d of the innovation with the 99%-statistically significant values of s_d, k_d and J_d marked in bold; (d) Heteroscedasticity measures h_{1b}, h_{1b} and h_{4b}, of σ_{yb}; (e) Maximum variance fraction β_g of the Gaussian error; (f) Bounds of the tuning factors: f_{o1}, f_{b1}, f_{o2} and f_{b2} of the specified standard deviation of errors. See text for details concerning quantities in (c), (d), (e) and (f).

Sample	4s	5s	6s	7s	11s	12s	14s	15s	5i	6i	7i	14i	15i
N	5907	4419	4132	3580	4426	6741	4081	4234	4980	4306	2043	3237	3581
σ_{yo}	0.6	0.6	0.6	0.75	1.0	2.0	0.5	0.6	0.6	0.6	0.75	0.5	0.6
$m(\sigma_{yb})$	0.083	0.126	0.243	0.432	0.904	1.040	0.354	0.207	0.378	1.031	2.502	1.442	0.618
$\mathrm{std}(\sigma_{yb})$	0.013	0.027	0.058	0.132	0.252	0.425	0.135	0.050	0.059	0.186	0.419	0.256	0.111
b_d	−0.01	−0.08	−0.13	−0.06	−0.02	0.17	−0.07	−0.04	0.03	−0.05	0.03	0.03	−0.01
σ_d	0.184	0.221	0.343	0.422	1.011	1.305	0.321	0.274	0.261	0.439	0.502	0.448	0.374
s_d	**−0.70**	**−0.63**	**−0.25**	0.10	**0.38**	**0.27**	−0.01	**−0.22**	**−0.34**	**−0.21**	**−0.26**	**−0.14**	**−0.21**
k_d	**1.02**	**0.90**	**0.35**	**0.56**	**0.65**	**2.07**	0.14	**0.23**	−0.07	**−0.47**	**−0.24**	**−0.33**	**−0.28**
J_d	**0.063**	**0.050**	**0.008**	**0.007**	**0.021**	**0.095**	0.0004	**0.005**	**0.010**	**0.008**	**0.007**	**0.004**	**0.005**
h_{1b}	0.99	0.98	0.97	0.96	0.96	0.93	0.93	0.97	0.99	0.98	0.99	0.98	0.98
h_{3b}	1.04	1.08	1.09	1.19	1.11	1.23	1.29	1.10	1.03	1.04	1.04	1.04	1.05
h_{4b}	1.10	1.23	1.28	1.61	1.33	1.72	2.08	1.32	1.10	1.11	1.10	1.12	1.12
β_g	0.646	0.675	0.852	0.983	0.810	0.965	1.000	0.843	0.637	0.566	0.622	0.643	0.638
f_{o1}	0.248	0.304	0.528	0.557	0.910	0.641	0.643	0.419	0.347	0.551	0.528	0.719	0.499
f_{b1}	1.762	1.409	1.266	0.925	0.969	1.141	0.848	1.182	0.544	0.316	0.156	0.245	0.477
f_{o2}	0.184	0.211	0.220	0.074	0.441	0.123	0.014	0.181	0.262	0.482	0.412	0.536	0.376
f_{b2}	1.305	0.978	0.527	0.122	0.470	0.219	0.018	0.511	0.411	0.276	0.122	0.183	0.359

Conclusion

- Innovation is very distinctly non-gaussian

What has been done here. Determine probability distributions of observation and/or background error by Maximum Entropy Method, *i*. *e*. determine distributions that have maximum entropy while being compatible with observed 1- to 4-order moments of innovation.

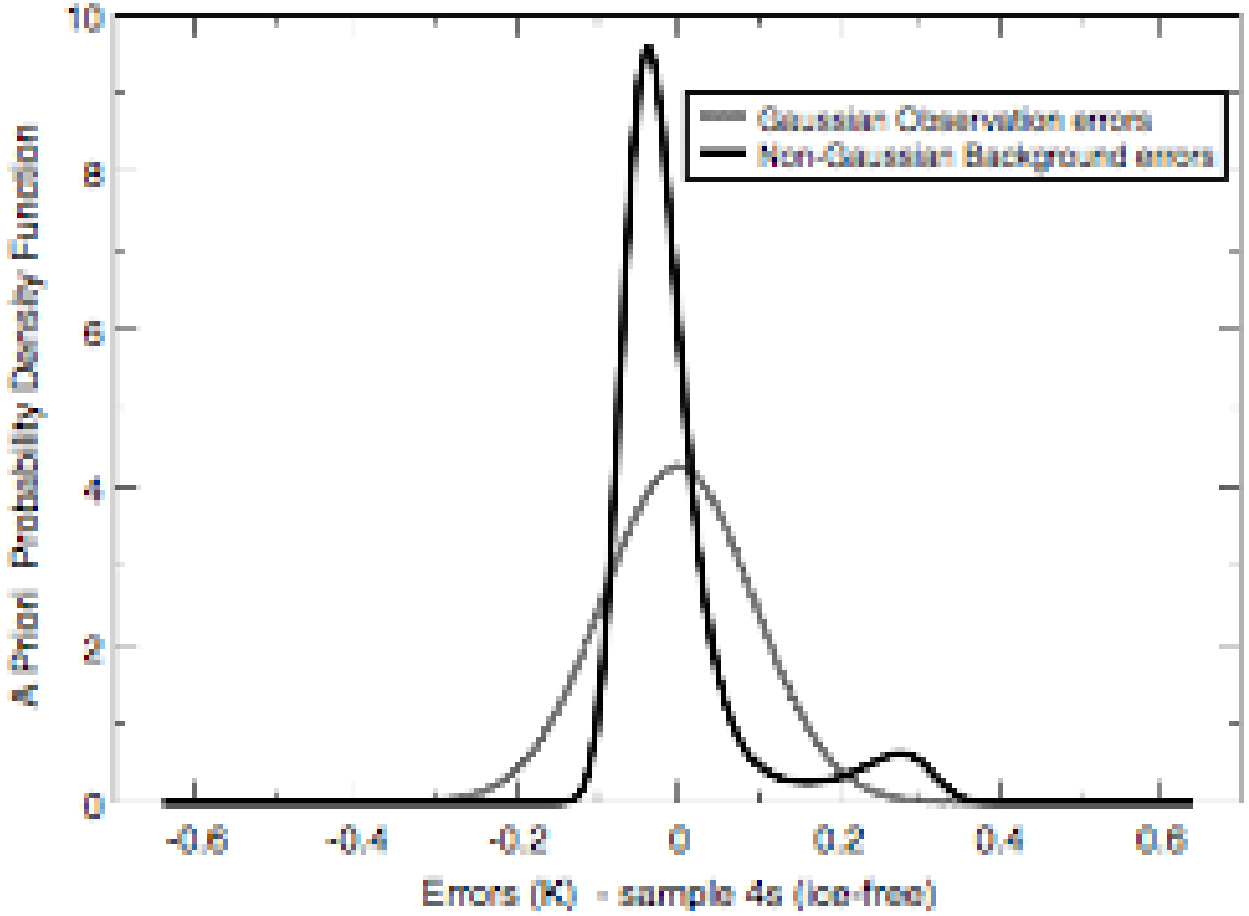
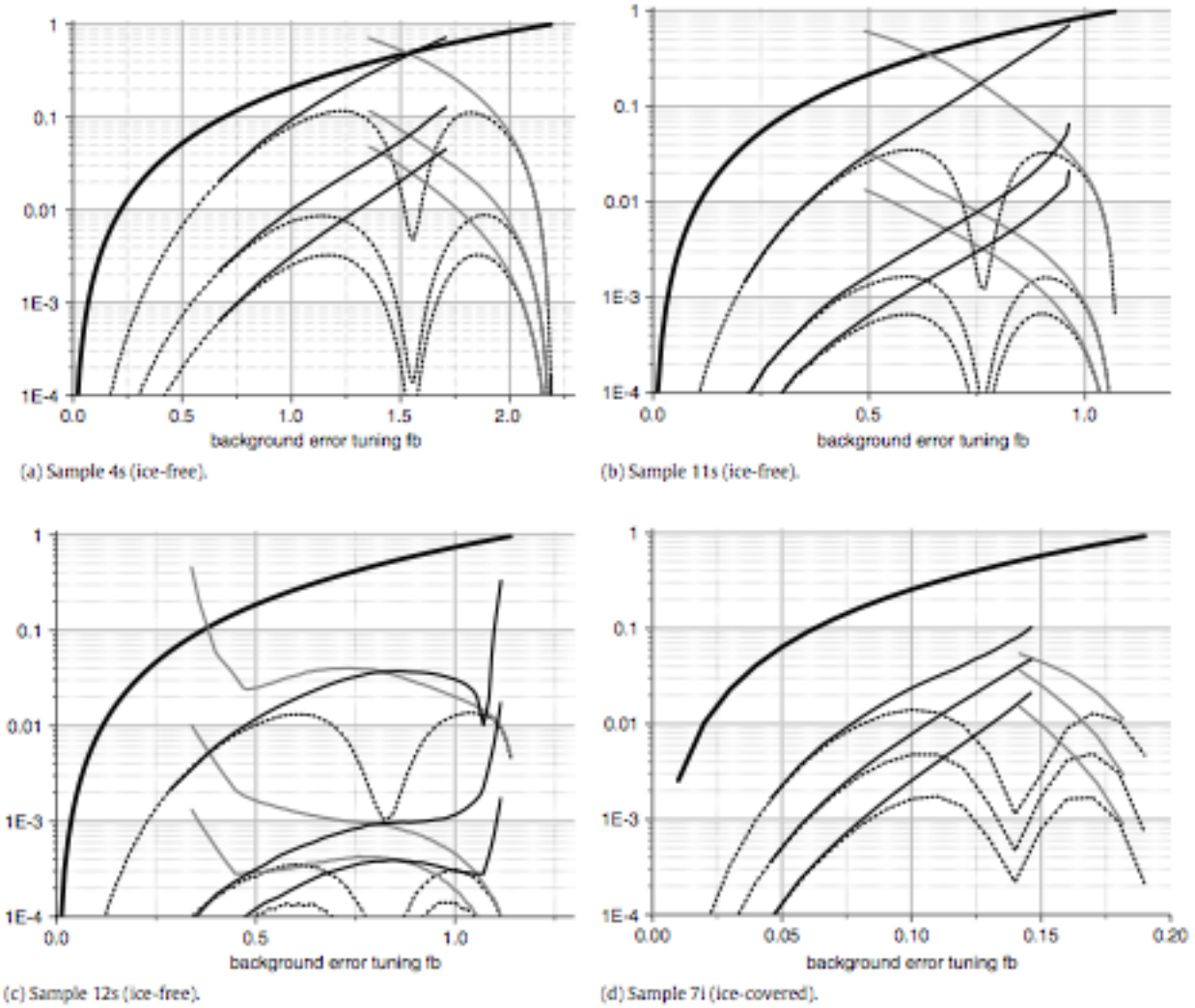


Fig. 5. *A priori* pdf of Gaussian observation errors (grey line) and non-Gaussian background errors (dark line), for sample 4s, found by the Maximum Entropy method, fixing the error variance partition: $\beta_b = 0.47$, $\beta_o = 0.53$ and $s_b = 2.18$, $k_b = 4.65$.



(a) Sample 4s (ice-free).

(b) Sample 11s (ice-free).

(c) Sample 12s (ice-free).

(d) Sample 7i (ice-covered).

Fig. 12. Common to all figures (a) to (d): Solid thick line: Value of β_b with the tuned mean variance of background errors. Three sets of curves with the scores MSC_j ($j = a, b, c$) and non-Gaussianity source scenarios: (1) Gaussian background errors: black thin solid lines; (2) Gaussian observation errors: grey thin solid lines and (3) *Maximum entropy sharing of* non-Gaussianity: black dotted lines. The values of the partial scores MSC_a, MSC_b, and MSC_c for each scenario of non-Gaussianity and the same tuning factor β_b obey in all cases to: $MSC_a < MSC_b < MSC_c$. (a) to (d) correspond respectively to samples 4s, 11s, 12s and 7i. The graphics are in logarithmic scale, values of the partial scores below 10^{-4} are not plotted. See text for details.

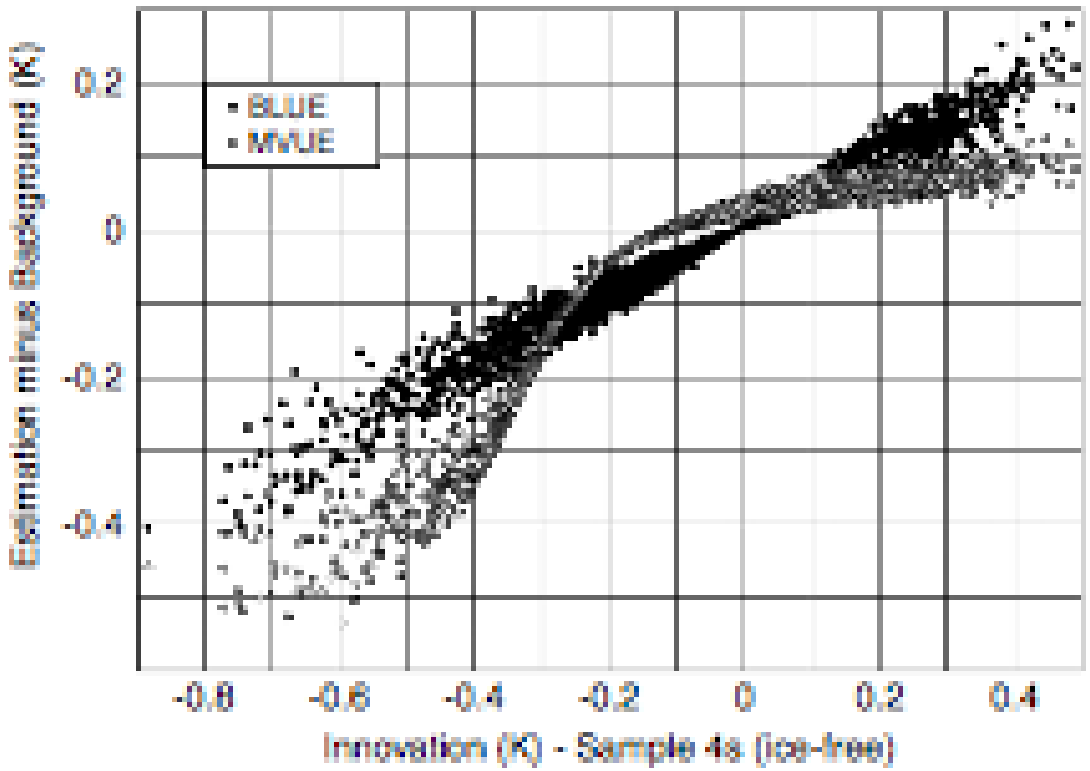


Fig. 13. Scatter plot of the background corrections: Δ_{Ub} (dark spots) and Δ_{Mb} (grey spots), respectively for the BLUE and the MBUE as function of the innovation for the sample 4s, fixing the error variance partition: $\beta_b = 0.47$, $\beta_o = 0.53$.