CERN IT Department

Information and treatment of unknown correlations in BLUE combinations

Andrea Valassi (CERN IT-ES)

Roberto Chierici (IPN Lyon)

Top LHC Working Group, 29th November 2012

Motivation

CERN IT Department

- **Best Linear Unbiased Estimators (BLUE) are widely used for precision measurement combinations**
  - e.g. top mass @ Tevatron and @ LHC
- **Larger event statistics have led (and will lead more and more) to combinations dominated by systematic errors**
  - And by the *correlations* between those systematics…
- **In parallel, negative combination coefficients are appearing (and may appear more and more) in BLUE averages**
  - We should not confuse these coefficients ("weights" in BLUE weighted central values) with the statistical impact (another type of "weight") of a measurement
- **Answer to typical questions**
  - How can we assess the "relative importance" of a measurement ("how much is my experiment/measurement contributing to a combination")?
  - What is the scientific interpretation of negative BLUE coefficients ("are we correctly estimating the correlations used as inputs to the combination")?

Main messages in this talk

1. **Absolute values of (negative) BLUE coefficients should not be used to assess the "relative importance" of measurements**
   – We suggest an alternative procedure based on Fisher's information

2. **Negative BLUE coefficients indicate a "high correlation" regime**
   – We show that this is true in general for N measurements of one observable

3. **Correlation estimates in this regime are not at all "conservative" and may need to be critically reassessed**
   – Overestimated correlations may lead to largely underestimated combined systematics and total errors
   – We propose a few tools to help in this critical review of correlations (derivatives of information, information minimization, covariance "onionization", marginal inflow of information)
   – **But the ideal solution is still to measure correlations in data and MC**

Outline

- **Part 1 – information in BLUE**
  - "Relative importance" of the measurements in the presence of correlations
- **Part 2 – interpretation of negative BLUE coefficients**
  - "Low-correlation" and "high-correlation" regimes
- **Part 3 – "conservative" estimates of correlations?**
  - The risks in overestimating correlations and a few hints to try and avoid this
- **Conclusions**

Part I :
information in BLUE

- **"Relative importance" of individual measurements with correlations**
  - BLUE primer
  - *What is the "contribution to knowledge" from various sources*
    - Absolute values of BLUE coefficients are used in the TOPLHCWG note
    - The problem with absolute values of BLUE weights
  - Fisher's information
    - A proposed new way for presenting combinations: information weights
    - First look at TOP LHC WG combination

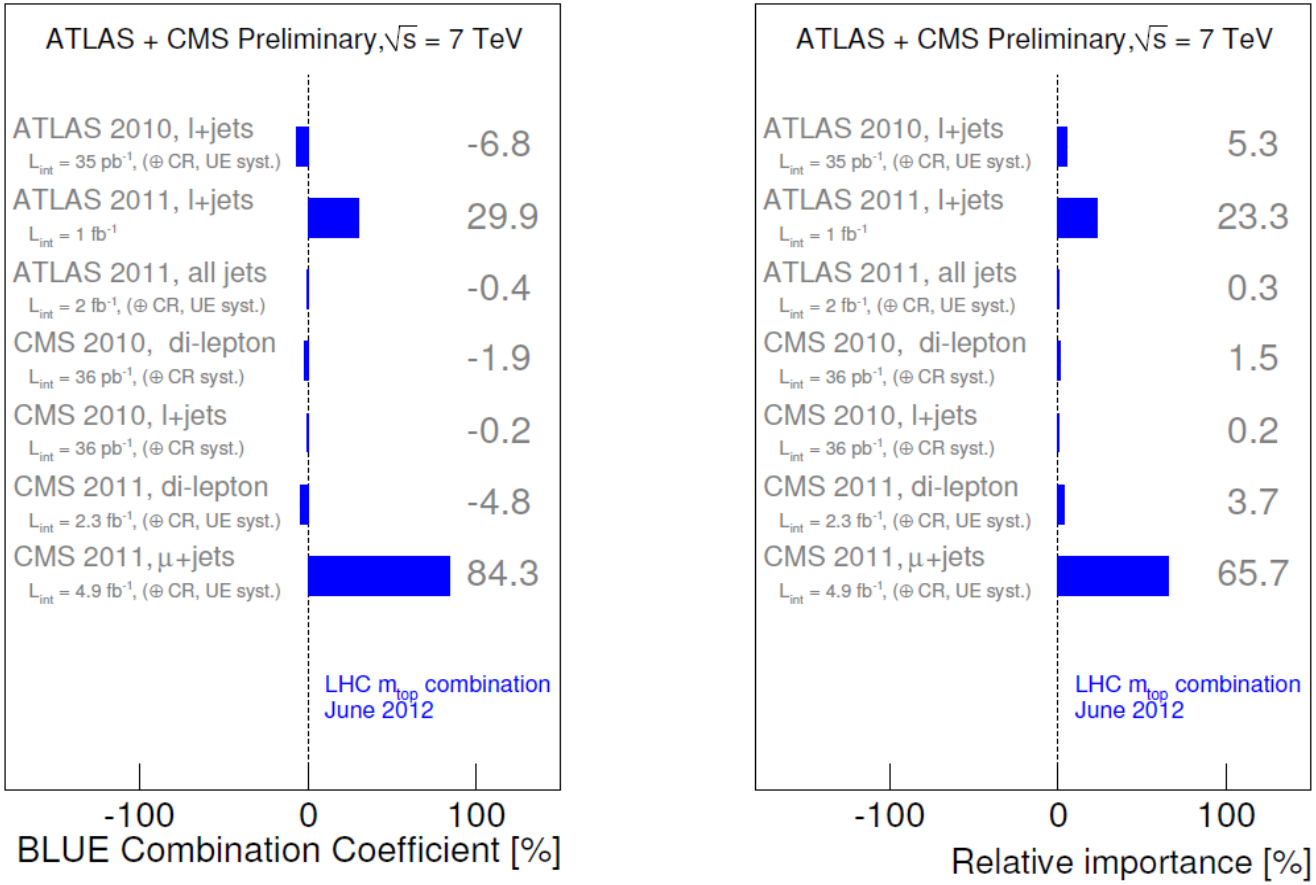
BLUE primer

CERN IT Department

- **Simple case *[Lyons]* : measurements y_i of a single observable Y**
  - A *linear* estimator $\hat{Y} = \tilde{\boldsymbol{\lambda}}\mathbf{y} = \sum_{i=1}^{n} \lambda_i \, y_i$ is *unbiased* if $\tilde{\mathbf{U}}\boldsymbol{\lambda} = \sum_{i=1}^{n} \lambda_i = 1$
    [where U is a vector of all 1's, i.e. contraction by U means summing all indexes]
  - The square of the error on the combined estimate Ŷ is $\mathrm{var}(\hat{Y}) = \sigma_{\hat{Y}}^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \, \mathcal{M}_{ij} \, \lambda_j$
    [where M is the (NxN) input covariance of the N measurements]
    NB Assume Gaussian pdf with M known a priori and independent of unknown parameter
  - The <u>*best* linear unbiased estimate</u> is given by $\lambda_i = \frac{(\mathcal{M}^{-1}\mathbf{U})_i}{(\tilde{\mathbf{U}}\mathcal{M}^{-1}\mathbf{U})}$
    (i.e. the error is minimized for)
  - *The coefficients λ_i in the BLUE linear combination can be negative*
    [a.k.a. "weights" in the BLUE weighted average]
  - The combined error on the BLUE is equal to: $\mathrm{var}(\hat{Y}) = \sigma_{\hat{Y}}^2 = \frac{1}{(\tilde{\mathbf{U}}\mathcal{M}^{-1}\mathbf{U})}$

Negative weights (and absolute weights) @LHC

CERN IT Department



http://cdsweb.cern.ch/record/1460097

TOPLHC NOTE

ATLAS-CONF-2012-095
CMS PAS TOP-12-001

June 2012

Figure 1: (a): Input measurements and result of the LHC combination (see also Table 4); (b, c) : BLUE combination coefficients and relative importance of the input measurements. The relative importance of a measurement is defined as $|w_i|/\sum_i |w_i|$, where $|w_i|$ is the absolute value of the BLUE combination coefficient for the i^{th} input measurement.

The problem with absolute values

CERN IT Department

- **Take A, B, C, with *C uncorrelated to A or B*:** $\left(\begin{array}{cc|c} \sigma_A^2 & \rho_{AB}\sigma_A\sigma_B & 0 \\ \rho_{AB}\sigma_A\sigma_B & \sigma_B^2 & 0 \\ \hline 0 & 0 & \sigma_C^2 \end{array}\right)$

- **Combining (A,B,C) in one go gives the same result for the combined error and for the linear coefficient of C as first combining (A,B) and then adding C:**

$$\frac{1}{(\sigma_{\hat{Y}}^2)_{ABC}} = \frac{1}{(\sigma_{\hat{Y}}^2)_{AB}} + \frac{1}{\sigma_C^2} \qquad (\lambda_C)_{ABC} = \frac{1/\sigma_C^2}{1/(\sigma_{\hat{Y}}^2)_{AB} + 1/\sigma_C^2}$$

- **The "relative importance" (RI) of C based on absolute values in combining (A,B,C) in one go is** $(\mathrm{RI}_C)_{A,B,C} = \frac{|(\lambda_C)_{ABC}|}{|(\lambda_A)_{ABC}| + |(\lambda_B)_{ABC}| + |(\lambda_C)_{ABC}|}$

- **It can easily be shown that there are two problems with this definition if $\lambda_B<0$ (which happens if $\rho_{AB}>\sigma_A/\sigma_B$ with $\sigma_A<\sigma_B$ – see later in this talk):**
  – first, RI of C is different if C is combined with A and B separately(lower) or if it is combined with the combination of A and B (higher)
  – second, RI of C underestimates the contribution of C with respect to its more intuitive (and "correct"?) assessment $\frac{1/\sigma_C^2}{1/(\sigma_{\hat{Y}}^2)_{AB} + 1/\sigma_C^2} = (\lambda_C)_{ABC}$

$\rightarrow$ *the "information weight" we propose for C is exactly equal to* $(\lambda_C)_{ABC}$

Fisher's information primer

- **Given a pdf P(y,X) for the N measurements y_i and the n parameters X_α, the (nxn) information matrix $I_{\alpha\beta}$ is defined as the covariance of $\partial \log P / \partial X_\alpha$**

- **Information has many nice properties [James]:**
  - Information is additive for independent measurements
  - Information $I_{\alpha\beta}$ is defined with respect to the set of parameters X_α to be estimated
  - Information is related to precision: the greater the available information, the more precise the estimates can be – the inverse of the information matrix is a lower bound (Cramer-Rao) for the covariance of any estimates of the parameters X_α
  - Information is positive semi-definite (for one observable it is a scalar ≥ 0)

- **Assuming P is multivariate Gaussian, the covariance of the BLUE estimates is equal to the inverse of the information matrix**
  - For a single observable (e.g. m_{top}), information is the inverse of the combined variance

$$I^{(Y)} = \tilde{\mathbf{U}}\mathcal{M}^{-1}\mathbf{U} = \frac{1}{\mathrm{var}(\hat{Y})} = \frac{1}{\sigma^2_{\hat{Y}}}$$

- **For all practical purposes, in the following, "information" about one parameter is just a synonym for the inverse of its variance**
  - WE WILL ALWAYS BE TALKING ABOUT A SINGLE OBSERVABLE IN WHAT FOLLOWS

Our proposal: Information Weights

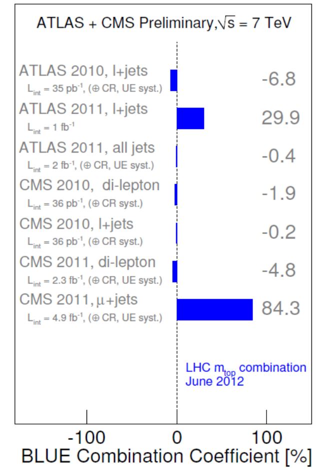
CERN IT Department

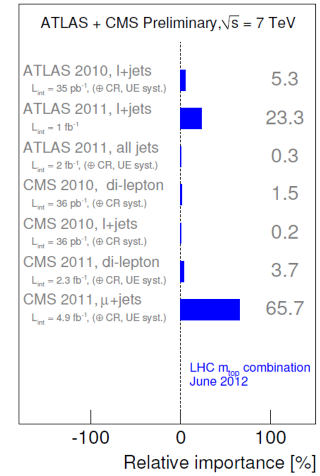
- **It is not only individual measurements, but also their correlation, that contribute to the information available in a combination**
  - The information contributed by correlations comes from the *collective interplay of the measurements* and cannot (in general) be attributed to any of them individually
  - We propose to distinguish between **the information weights from the measurements** and **the information weight from the ensemble of correlations**

- **Our definition of Information Weights (IW)**
  - Information weight for each measurement (always > 0): $$\mathrm{IW}_i = \frac{1/\sigma_i^2}{1/\sigma_{\hat{Y}}^2}$$
    - In the absence of any correlations, this coincides with the BLUE linear coefficient λ
    - Also true if one measurement is uncorrelated (e.g. C in previous A,B,C example)
  - Information weight for the correlations is all the rest (may be <0): $$\mathrm{IW}_{\mathrm{corr}} = \frac{1/\sigma_{\hat{Y}}^2 - \sum_i 1/\sigma_i^2}{1/\sigma_{\hat{Y}}^2}$$
    - In the absence of any correlations, this is equal to 0
  - For N measurements of 1 observable: # information weights are N (measurements) plus 1 (correlations)

- **One issue: not easy/obvious how to subdivide the correlation weight**
  - e.g. split up contributions by error source and/or by pair of measurements?
  - the interplay of correlations cannot be neglected, their effect is largely a collective one
  - will come back to this only at the very end

A practical example (more later): LHC m_{top}





	m_{top} [GeV]	ΔI [1/GeV^2]	λ	$\frac{\lvert\lambda\rvert}{\sum\lvert\lambda\rvert}$	$\frac{\Delta I}{I}$
CMS11μj	172.64 ± 1.53	0.425	84.3%	65.7%	83.8%
ATL11lj	174.53 ± 2.39	0.175	29.9%	23.3%	34.5%
ATL10lj	169.33 ± 6.32	0.025	-6.8%	5.3%	4.9%
CMS11ll	173.30 ± 2.96	0.114	-4.8%	3.7%	22.5%
CMS10ll	175.50 ± 6.49	0.024	-1.9%	1.5%	4.7%
ATL11aj	174.90 ± 4.44	0.051	-0.4%	0.3%	10.0%
CMS10lj	173.10 ± 3.41	0.086	-0.2%	0.2%	16.9%
Correlations	—	-0.392	—	—	-77.4%
Total	173.33 ± 1.40	0.507	100.0%	100.0%	100.0%

The real challenge (and the interesting part) is splitting up the correlation contribution and understanding where it comes from!

Part II :
negative BLUE coefficients

- **"Low-correlation" and "High-correlation" regimes**
  - The simple case of two measurements of one observable [Lyons]
    - Dependencies of BLUE coefficients and combined error and IW as the correlation varies
    - One linear coefficient λ goes to 0 and flips sign where the combined error is minimum
    - Interpretation of information in the two regimes λ>0 and λ<0
  - Generalization to the case of N measurements of one observable
    - Information inflow – adding the N^{th} measurement with its N-1 correlations
    - Information derivatives – analyzing Nx(N-1)/2 correlations independently
    - **Low-correlation regime (all λ>0) and high-correlation regime (one or more λ<0)**

WE WILL ONLY DISCUSS POSITIVE CORRELATIONS IN THE FOLLOWING

2 measurements of 1 observable

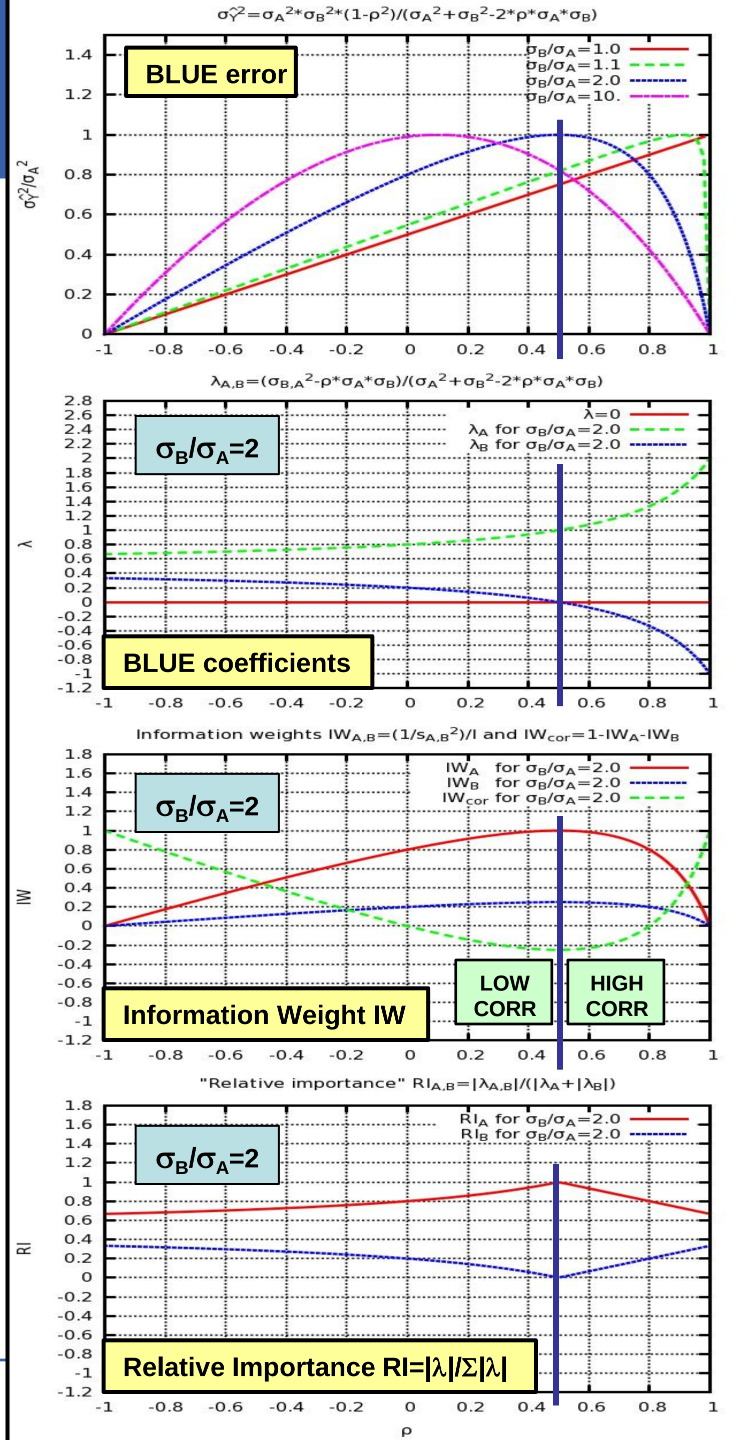
- **For two measurements A and B with errors $\sigma_A < \sigma_B$ and correlation ρ *[Lyons et al]* :**

$$\hat{Y} = \lambda_A y_A + \lambda_B y_B \qquad \lambda_A = \frac{\sigma_B^2 - \rho\sigma_A\sigma_B}{\sigma_A^2 + \sigma_B^2 - 2\rho\sigma_A\sigma_B}$$

$$\sigma_{\hat{Y}}^2 = \frac{\sigma_A^2\sigma_B^2(1-\rho^2)}{\sigma_A^2 + \sigma_B^2 - 2\rho\sigma_A\sigma_B} = \frac{1}{I} \qquad \lambda_B = \frac{\sigma_A^2 - \rho\sigma_A\sigma_B}{\sigma_A^2 + \sigma_B^2 - 2\rho\sigma_A\sigma_B}$$

- **Varying ρ, the combined error is maximum (equal to σ_A) for $\rho=\sigma_A/\sigma_B$, where λ_B flips sign**
  - ***Low correlation region – $\rho<\sigma_A/\sigma_B$***
    - Both λ_A and λ_B are > 0
    - Info decreases (error increases) as ρ increases
  - ***High correlation region – $\rho>\sigma_A/\sigma_B$***
    - One coefficient λ_B is < 0
    - Info increases (error decreases) as ρ increases
  - ***Boundary (most conservative corr.) – $\rho=\sigma_A/\sigma_B$***
    - One coefficient λ_B is = 0
    - Error is maximum = σ_A (marginal info from B is 0)

- **Three information weights (A, B, correlation)**
  - Sum is equal to 1 by construction
  - IW from ρ is minimum (largest negative) for λ_B=0



Two measurements – interpretation

CERN IT Department

- **Low correlation regime, $\lambda_B>0$ (i.e. $\sigma_A^2 - \rho\sigma_A\sigma_B > 0$)**
  – The covariance can be seen as the sum of a common error and an uncorrelated error

>0 (positive definite matrix)

$$\begin{pmatrix} \sigma_A^2 & \rho\sigma_A\sigma_B \\ \rho\sigma_A\sigma_B & \sigma_B^2 \end{pmatrix} = \begin{pmatrix} \rho\sigma_A\sigma_B & \rho\sigma_A\sigma_B \\ \rho\sigma_A\sigma_B & \rho\sigma_A\sigma_B \end{pmatrix}_{\mathrm{com}} + \begin{pmatrix} \sigma_A^2 - \rho\sigma_A\sigma_B & 0 \\ 0 & \sigma_B^2 - \rho\sigma_A\sigma_B \end{pmatrix}_{\mathrm{unc}}$$

  – You can combine based on the uncorrelated error (compute statistical weights from basic error propagation) and add the common error only at the end
    - *Remember: this happens if the off-diagonal covariance $\rho\sigma_A\sigma_B$ is smaller than the smaller of the two variances σ_A^2 (this will be used later in our "onionization" prescription proposal)*
  – **Adding the less precise measurement B to the combination brings in additional information because B contributes independent (uncorrelated) knowledge about the unknown parameter**

- **High correlation regime, $\lambda_B<0$ (i.e. $\sigma_A^2 - \rho\sigma_A\sigma_B < 0$)**
  – *The covariance can NOT be seen as the sum of a common error and an uncorrelated error*
  – **Adding the less precise measurement B to the combination brings in additional information because B helps to constrain a correlated error on A to which it has a different sensitivity**
    - ρ may be overestimated or this may be perfectly legitimate: *you may leverage on the different sensitivity of two measurements to a common background or a common MC parameter to reduce the uncertainty on them*
    - The additional information only comes from the interplay of B with A – neither can claim it as their exclusive merit

- **Boundary is $\lambda_B=0$ – where B brings no additional information**
  – **Adding the less precise measurement B to the combination brings in NO additional information because its error has a part common with A plus an additional "uncorrelated" component**

Generalization to N measurements

(REMINDER: WE WILL ONLY DISCUSS POSITIVE CORRELATIONS IN THE FOLLOWING)

- **The previous discussion can be generalized from 2 to N measurements**
  - Fixing the variances on the measurements and varying only the correlations between them, any BLUE combination can fall in one of two regimes:
    - ***EITHER Increasing any correlation decreases information*** and increases the combined error ("low" correlation regime because it includes the case where correlations are all 0)
    - ***OR Increasing some correlations increases information*** and decreases the combined error ("high" correlation regime because you must increase a correlation beyond a threshold to enter it)

- **It can be shown that these regimes are defined by the BLUE coefficients:**
  - **Low correlations** **(info decreases if any correlation increases)** ↔ **All BLUE coefficients are ≥0**
  - **High correlations** **(info increases if a correlation increases)** ↔ **Some BLUE coefficients are <0**
  - In other words: if any BLUE coefficients are <0, it is because some correlations are high

- **Marginal info from N^th measurement: fix all variances, vary N-1 correlations 'c'**

$$\mathcal{M}_{N\times N} = \begin{pmatrix} \mathcal{D}_{(N-1)\times(N-1)} & \mathbf{c} \\ \tilde{\mathbf{c}} & d \end{pmatrix} \qquad \Delta I_N \propto \left((\tilde{\mathbf{u}}\mathcal{D}^{-1}\mathbf{c}) - 1\right)^2 \geq 0 \qquad \lambda_N \propto -\left((\tilde{\mathbf{u}}\mathcal{D}^{-1}\mathbf{c}) - 1\right)$$

- **General case: fix variances, vary Nx(N-1)/2 correlations and use derivatives**
  - Information is minimum when some coefficients are 0
  - If it is impossible to properly estimate correlations, the most "conservative" option is to exclude some measurements…

$$\frac{\partial I}{\partial \rho_{ij}} = -2I^2\lambda_i\lambda_j\sigma_i\sigma_j$$

Summary of part II

CERN IT Department

- **If any BLUE coefficients are <0, it is because some correlations are high**
  - The BLUE combination is in a "high correlation" regime
- **The contribution to information from measurements with BLUE coefficients <0 comes only from their interplay with other measurements through correlations**
  - They help constrain systematic uncertainties on other measurements
- **If any BLUE coefficients are <0, the total BLUE error is smaller than that which would be obtained if some correlations were smaller**
  - Are correlations really as high or are they being overestimated?

IF any BLUE coefficients are <0

AND IF correlations were "conservatively" estimated to be 100%,

THEN correlation estimates should be reassessed

See part III for a few tools and suggestions

PRELIMINARY!
WORK IN PROGRESS!

Part III : "conservative" estimates of correlations

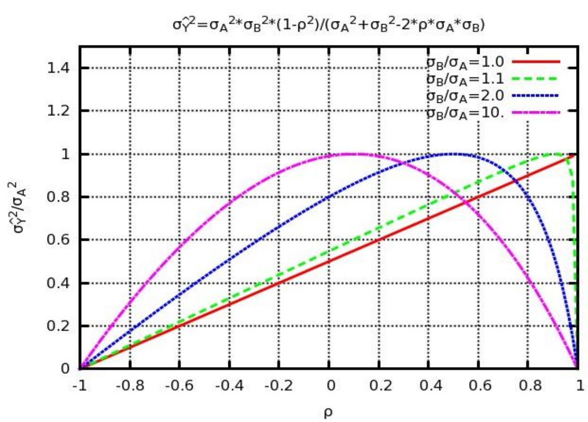
- **The risks in overestimating correlations and a few hints to try and avoid this – with practical examples from m_{top} @ LHC**
  - "100%" is not conservative when statistical errors are low
  - Using information derivatives to identify the most relevant correlations
  - Minimization and "onionization" procedures
  - Splitting up the correlation contribution to information using marginal inflow
  - Practical recommendations

WE WILL ONLY DISCUSS POSITIVE CORRELATIONS IN THE FOLLOWING

Is ρ=100% "conservative"?

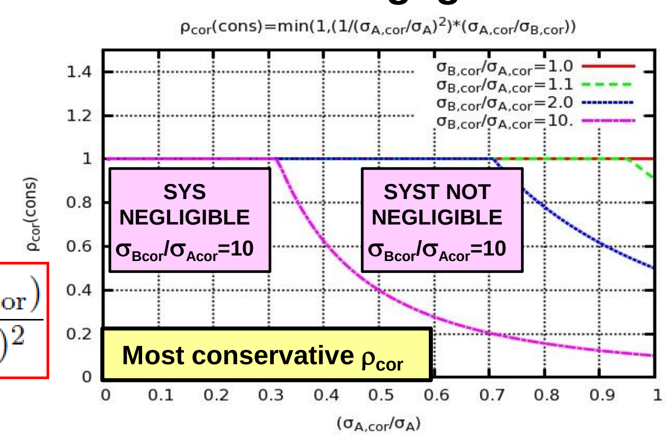
- **Two measurements AB, each with a statistical and a systematic error**
  - Statistical errors are uncorrelated, systematic errors are correlated with a correlation ρ_{cor}

- **If ρ_{cor} is unknown, we want to look for the most "conservative" of ρ_{cor} i.e. that which makes the combined BLUE error as large as possible**
  - The most conservative ρ_{cor} is that which makes the *total* correlation ρ as large as possible, as long as ρ remains in the "region low-correlation"$\rho<\sigma_A/\sigma_B$ (assuming $\sigma_A<\sigma_B$)
  - ***If measurements are statistically dominated, the most conservative ρ_{cor} is 100%***
  - **But the most conservative ρ_{cor} is < 100% if systematic errors are not negligible**



$$\rho = \frac{\rho_{\rm cor}\,\sigma_{A,{\rm cor}}\,\sigma_{B,{\rm cor}}}{\sigma_A \sigma_B} \le \frac{\sigma_A}{\sigma_B}$$

$$\rho_{\rm cor} \le \frac{\sigma_A^2}{\sigma_{A,{\rm cor}}\,\sigma_{B,{\rm cor}}} = \frac{(\sigma_{A,{\rm cor}}/\sigma_{B,{\rm cor}})}{(\sigma_{A,{\rm cor}}/\sigma_A)^2}$$



With N measurements, finding the most conservative correlations is not as easy!

Prioritizing correlations to review

- **So, I think I overestimated my correlations: and now, where do I start?**
  – ***1 - Which correlations should I consider reassessing first?***

- **Use information derivatives to analyze which correlations (between which 2 measurements and for which error sources) have the largest effect**
  – Rescale the nominal covariance using a different rescaling factor (between 0 and 1 – do not increase correlations or flip their sign) for each off-diagonal element and error source

$$\mathcal{M}_{ij}^{[s]} \rightarrow (\mathcal{M}')_{ij}^{[s]} = \begin{cases} f_{ij}^{[s]} \mathcal{M}_{ij}^{[s]} & \text{if } i \neq j, \\ \mathcal{M}_{ii}^{[s]} & \text{if } i = j, \end{cases} \qquad 0 \leq f_{ij}^{[s]} \leq 1$$

  – Will show in the following the derivatives (at nominal correlations) with respect to these scale factors

$$\frac{\partial I'}{\partial f_{ij}^{[s]}} = -2(I')^2 \lambda'_i \lambda'_j \mathcal{M}_{ij}^{[s]}$$

  – Will normalize derivatives to total information at nominal correlations to make them a-dimensional

- **For the top measurement at LHC there would be 21x16 derivatives**
  – 21 combinations (of 7 measurements) and 16 correlated error sources
  – They should be studied individually, but in the following will show them summed by off-diagonal element (21) and by error source (16); the grand-total is always the derivative by a global scale factor

LHC m_{top} information derivatives

	Stat	iJES	bJES	dJES	rJES	Lept	MC	Rad	CR	PDF	DTMO	UE	BGMC	BGDT	Meth	MHI	TOTAL
$\frac{1}{I}@1 \times \frac{\partial I}{\partial f}@0$	0.000	0.000	-0.879	-0.499	0.000	-0.001	-0.021	-0.813	-0.324	-0.034	-0.099	-0.154	-0.022	0.000	0.000	-0.054	-2.901
$\frac{1}{I}@1 \times \frac{\partial I}{\partial f}@1$	0.000	0.000	0.051	0.051	0.000	0.000	0.008	0.051	-0.032	0.002	0.019	0.035	0.011	0.000	0.000	0.022	0.219

1st line: derivatives (all <0 !) at 0 correlations
2nd line: derivatives (most >0 !) at nominal correlations
Both normalized by info at nominal correlations

PRELIMINARY!
WORK IN PROGRESS!

$\frac{1}{I}@1 \times \frac{\partial I}{\partial f}@0$	ATL10lj	ATL11lj	ATL11aj	CMS10ll	CMS10lj	CMS11ll	CMS11μj
ATL10lj	—	—	—	—	—	TOTAL	
ATL11lj	-0.164	—	—	—	—	-2.901	
ATL11aj	-0.070	-0.223	—	—	—	—	—
CMS10ll	-0.009	-0.026	-0.010	—	—	—	—
CMS10lj	-0.029	-0.098	-0.034	-0.058	—	—	—
CMS11ll	-0.034	-0.127	-0.045	-0.083	-0.259	—	—
CMS11μj	-0.107	-0.366	-0.125	-0.140	-0.376	-0.519	—

$\frac{1}{I}@1 \times \frac{\partial I}{\partial f}@1$	ATL10lj	ATL11lj	ATL11aj	CMS10ll	CMS10lj	CMS11ll	CMS11μj
ATL10lj	—	—	—	—	—	TOTAL	
ATL11lj	0.197	—	—	—	—	0.219	
ATL11aj	-0.004	0.007	—	—	—	—	—
CMS10ll	-0.005	0.009	0.000	—	—	—	—
CMS10lj	-0.001	0.001	0.000	0.000	—	—	—
CMS11ll	-0.010	0.023	0.000	-0.007	-0.001	—	—
CMS11μj	0.149	-0.318	0.005	0.058	0.004	0.111	—

	Global Factor
$\frac{1}{I}@1 \times \frac{\partial I}{\partial f}@0$	-2.901
$\frac{1}{I}@1 \times \frac{\partial I}{\partial f}@1$	0.219

- **The first correlations that should be reviewed (if needed) are:**
  - Those between ATL10lj/ATL11lj, ATL10lj/CMS11mj, CMS11mj/CMS11ll
  - Those for the bJES, dJES and RAD error sources
- **Next slides: look at dJES and bJES for ATL10lj/ATL11lj and CMS11mj/CMS11ll**

LHC m_{top} systematics

Uncertainty Categories			Size [GeV]							Correlation	
			ATLAS			CMS					
Tevatron	ATLAS	CMS	2010 l+jets	2011 l+jets	2011 all jets	2010 di-l	2010 l+jets	2011 di-l	2011 μ+jets	ρ_{exp}	ρ_{LHC}
Statistics			4.0	0.6	2.1	4.6	2.1	1.2	0.4	0	0
iJES	Jet Scale Factor	Jet Scale Factor		0.4					0.4	0	0
aJES											
bJES	JES_{b-jet}	JES_{b-jet}	2.5	1.6	1.4	0.9	0.9	1.1	0.7	1	0.5
cJES											
dJES	$JES_{light-jet}$	$JES_{light-jet}$	2.1	0.7	2.1	2.1	2.1	2.0	0.2	1	0
rJES		residual-JES				3.3				0	0
LepPt		Lepton p_T Scale				0.3		0.2		1	0
MC	MC Generator	MC Generator	0.7	0.3	0.5	0.4		0.1			
	Hadronisation		0.7	0.2	(*)						
	Sum	Sum	1.0	0.4	0.5	0.4		0.1		1	0.5
Rad	ISR/FSR	ISR/FSR	2.5	1.0	1.7	0.2	0.2				
		Q-Scale				0.6	1.1	0.4	0.8		
		Jet-Parton Scale				0.7	0.4	0.7	0.3		
	Sum	Sum	2.5	1.0	1.7	0.9	1.2	0.8	0.8	1	0.5
CR	Colour Recon.		*0.6*	0.6	*0.6*	*0.5*	*0.5*	0.5	0.5	1	1
PDF	Proton PDF	Proton PDF	0.5	0.1	0.6	0.5	0.1	0.4	0.1	1	1
	Jet Energy Res.	Jet Energy Res.	0.9	0.1	0.3	0.5	0.1	0.3	0.2		
	Jet Rec. Eff.		0.5	< 0.05	0.2						
	b-tagging	b-tagging	0.5	0.3	0.3	0.4	0.1	0.5	0.2		
	E_T^{miss}	E_T^{miss}		0.1		0.1	0.4	0.4	0.1		
DetMod	Sum	Sum	1.2	0.3	0.5	0.7	0.4	0.7	0.3	1	0
UE	Underlying Event	Underlying Event	*0.6*	0.6	*0.6*	1.4	0.2	*0.6*	*0.6*	1	0
	W+jet Norm.		1.6								
	W+jet Shape		0.8	0.1							
		background				0.1	0.2		0.1		
BGMC	Sum	Sum	1.8	0.1		0.1	0.2		0.1	1	1
	W+jet Norm.			0.4							
	QCD Norm.	QCD Norm.	0.5	0.2			0.4	0.4			
	QCD Shape		0.4	0.3	1.9						
BGData	Sum	Sum	0.6	0.5	1.9		0.4	0.4		0	0
Method	Method Calib.	Method Calib.	0.4	0.1	1.0	0.3	0.1	0.4	0.2	0	0
MHI	Pile-up	Pile-up	0.7	< 0.05		1.0	0.1	0.2	0.4	1	1

Example:

- bJES and dJES are 100% correlated between ATL11lj and ATL10lj, or CMS11mj and CMS11ll, but have very **different variances** in the 2 measurements in each exp.

- The larger sensitivities of ATL10lj and CMS11ll to bJES and dJES with respect to ATL11lj and CMS11mj makes the former (overall less precise) measurements contribute information and reduce the combined error by **constraining the dJES and bJES systematics** in the latter (overall more precise) measurements

LHC m_{top} marginal combinations

CERN IT Department

	Comb. $m_{\rm top}$	Stat	iJES	aJES	bJES	cJES	dJES	rJES	Lept	MC	Rad	CR	PDF	DTMO	UE	BGMC	BGDT	Meth	MHI	χ^2/ndof
CMS11μj	172.64 ± 1.53	0.37	0.43	0.00	0.66	0.00	0.23	0.00	0.00	0.00	0.80	0.54	0.05	0.28	0.59	0.09	0.00	0.15	0.38	—
⊕ ATL11lj	173.02 ± 1.46	0.32	0.36	0.00	0.74	0.00	0.23	0.00	0.00	0.07	0.76	0.54	0.06	0.24	0.49	0.10	0.10	0.12	0.31	0.64/1
⊕ ATL10lj	173.44 ± 1.42	0.44	0.36	0.00	0.71	0.00	0.18	0.00	0.00	0.04	0.69	0.54	0.03	0.22	0.48	0.02	0.16	0.12	0.25	2.06/2
⊕ CMS11ll	173.40 ± 1.41	0.46	0.38	0.00	0.68	0.00	0.09	0.00	0.01	0.04	0.69	0.55	0.01	0.19	0.48	0.01	0.16	0.13	0.26	2.14/3
⊕ CMS10ll	173.34 ± 1.40	0.47	0.38	0.00	0.68	0.00	0.08	0.06	0.01	0.04	0.69	0.55	0.01	0.19	0.46	0.01	0.16	0.13	0.25	2.40/4
⊕ ATL11aj	173.33 ± 1.40	0.47	0.38	0.00	0.68	0.00	0.07	0.06	0.01	0.04	0.69	0.55	0.01	0.19	0.46	0.01	0.16	0.13	0.25	2.49/5
⊕ CMS10lj	173.33 ± 1.40	0.47	0.38	0.00	0.68	0.00	0.07	0.06	0.01	0.04	0.69	0.55	0.01	0.19	0.47	0.01	0.16	0.13	0.25	2.49/6

Table 8: LHC $m_{\rm top}$ BLUE combinations with the marginal addition of only one measurement at a time. The measurements are added in the following order: first those with $\lambda > 0$, ordered by decreasing λ; then those with $\lambda < 0$, ordered by decreasing $|\lambda|$.

- **The large effect of bJES and dJES correlations can also be seen by adding measurements one by one (order by decreasing λ>0 and then decreasing |λ|)**
  - The first 2 measurements (λ>0) bring uncorrelated information (reduce statistical errors)
  - The next 2 (λ<0) reduce common systematics because of high correlations
    - *The third measurement (λ = -7%) has a small effect on the combined error, but it reduces systematics (bJES, dJES, RAD…) and pushes the statistical error much higher! It also shifts the central value significantly. Is this reasonable or are correlations overestimated?*
  - The last 3 (λ<0, |λ|<2%) have virtually no effect on the combination

- **Note that (in general) all measurements contribute to χ^2 even if λ≤0**
  - It is not true that a measurement with λ=0 is effectively ignored
  - *Why does χ^2 not change here for a growing number of degrees of freedom?*

PRELIMINARY!
WORK IN PROGRESS!

Making correlations more conservative?

CERN IT Department

- **So, I think I overestimated my correlations: and now, where do I start?**
  - ***2 – Is there a way to (blindly) make correlations more conservative?***
- **We tested four procedures to reduce the nominal correlations**
  - *Minimize information by varying one scale factor per error source (16 for m_{top})*
  - *Minimize information by varying one scale factor per measurement pair (21 for m_{top})*
    - This may lead into non-physical space (non positive definite covariances) – give up in that case
  - *Minimize information by varying one global scale factor (1 for m_{top})*
  - *"Onionize" covariance matrices* as described in the next slide

But carefully measuring correlations is much better than any rule of thumb!

The "onionization" prescription

- **Generalize to N measurements an observation made for two: keep each off-diagonal covariance $\rho_{ij}\sigma_i\sigma_j$ smaller than both variances σ_i^2 and σ_j^2**
  - We do this for each error source of uncertainty [s]

$$\rho^{[s]}_{\mathrm{cor}}\sigma^{[s]}_{A,\mathrm{cor}}\sigma^{[s]}_{B,\mathrm{cor}} \le (\sigma^{[s]}_{A,\mathrm{cor}})^2 \quad \forall s$$
$$\rho^{[s]}_{\mathrm{cor}}\sigma^{[s]}_{A,\mathrm{cor}}\sigma^{[s]}_{B,\mathrm{cor}} \le (\sigma^{[s]}_{B,\mathrm{cor}})^2 \quad \forall s$$

$$\begin{pmatrix} (\sigma^{[s]}_{A,\mathrm{cor}})^2 & (\sigma^{[s]}_{A,\mathrm{cor}})^2 & (\sigma^{[s]}_{A,\mathrm{cor}})^2 & (\sigma^{[s]}_{A,\mathrm{cor}})^2 & \dots \\ (\sigma^{[s]}_{A,\mathrm{cor}})^2 & (\sigma^{[s]}_{B,\mathrm{cor}})^2 & (\sigma^{[s]}_{B,\mathrm{cor}})^2 & (\sigma^{[s]}_{B,\mathrm{cor}})^2 & \dots \\ (\sigma^{[s]}_{A,\mathrm{cor}})^2 & (\sigma^{[s]}_{B,\mathrm{cor}})^2 & (\sigma^{[s]}_{C,\mathrm{cor}})^2 & (\sigma^{[s]}_{C,\mathrm{cor}})^2 & \dots \\ (\sigma^{[s]}_{A,\mathrm{cor}})^2 & (\sigma^{[s]}_{B,\mathrm{cor}})^2 & (\sigma^{[s]}_{C,\mathrm{cor}})^2 & (\sigma^{[s]}_{D,\mathrm{cor}})^2 & \dots \\ \dots & \dots & \dots & \dots & \dots \end{pmatrix}$$
$$\sigma_{\mathrm{A,cor}} \le \sigma_{\mathrm{B,cor}} \le \sigma_{\mathrm{C,cor}} \le \sigma_{\mathrm{D,cor}}$$

  - This may lead to underestimated ρ and is not even enough to ensure "conservativeness"
    - But it may help avoid strange effects with highly correlated errors of very different sizes

PRELIMINARY! WORK IN PROGRESS!

- **Example: covariance for the RAD uncertainty**
  - e.g. reduce the covariance for ATL10lj and ATL11lj (very different variances: 2.5 and 1.0)

CMS11ll	0.59	0.61	0.73	0.39	0.91	0.65	0.94
CMS11μj	0.61	0.64	0.75	0.40	0.95	0.68	0.98
CMS10ll	0.73	0.75	0.89	0.48	1.12	0.80	1.15
ATL11lj	0.39	0.40	0.48	1.02	0.60	1.72	2.47
CMS10lj	0.91	0.95	1.12	0.60	1.41	1.01	1.45
ATL11aj	0.65	0.68	0.80	1.72	1.01	2.89	4.16
ATL10lj	0.94	0.98	1.15	2.47	1.45	4.16	5.98

Table 13: Onionization of the RAD covariance. Nominal RAD.

OLD

CMS11ll	0.59	0.59	0.59	0.59	0.59	0.59	0.59
CMS11μj	0.59	0.64	0.64	0.64	0.64	0.64	0.64
CMS10ll	0.59	0.64	0.89	0.89	0.89	0.89	0.89
ATL11lj	0.59	0.64	0.89	1.02	1.02	1.02	1.02
CMS10lj	0.59	0.64	0.89	1.02	1.41	1.41	1.41
ATL11aj	0.59	0.64	0.89	1.02	1.41	2.89	2.89
ATL10lj	0.59	0.64	0.89	1.02	1.41	2.89	5.98

Table 14: Onionization of the RAD covariance. Onion upper bound.

MAX

CMS11ll	0.59	0.59	0.59	0.39	0.59	0.59	0.59
CMS11μj	0.59	0.64	0.64	0.40	0.64	0.64	0.64
CMS10ll	0.59	0.64	0.89	0.48	0.89	0.80	0.89
ATL11lj	0.39	0.40	0.48	1.02	0.60	1.02	1.02
CMS10lj	0.59	0.64	0.89	0.60	1.41	1.01	1.41
ATL11aj	0.59	0.64	0.80	1.02	1.01	2.89	2.89
ATL10lj	0.59	0.64	0.89	1.02	1.41	2.89	5.98

Table 15: Onionization of the RAD covariance. Onionized RAD.

NEW!

LHC m_{top} – conservative correlations?

	Comb. m_{top}	Stat	iJES	bJES	dJES	rJES	Lept	MC	Rad	CR	PDF	DTMO	UE	BGMC	BGDT	Meth	MHI	χ^2/ndof
JUNE 2012 TOP WG	173.33 ± 1.40	0.47	0.38	0.68	0.07	0.06	0.01	0.04	0.69	0.55	0.01	0.19	0.47	0.01	0.16	0.13	0.25	2.49/6
MinimizeGlobalFactor1	173.29 ± 1.41	0.40	0.36	0.71	0.19	0.05	0.01	0.06	0.71	0.53	0.04	0.22	0.47	0.04	0.14	0.12	0.26	2.11/6
MinimizeErrorSrc18	173.22 ± 1.44	0.35	0.35	0.72	0.25	0.02	0.00	0.09	0.73	0.54	0.05	0.24	0.48	0.10	0.12	0.12	0.29	1.97/6
Onionize	173.08 ± 1.44	0.31	0.33	0.72	0.26	0.01	0.00	0.07	0.76	0.54	0.06	0.24	0.47	0.09	0.10	0.11	0.28	1.67/6
MinimizeOffDiagonal21	173.02 ± 1.46	0.32	0.36	0.74	0.23	0.00	0.00	0.07	0.76	0.54	0.06	0.24	0.49	0.10	0.10	0.12	0.31	1.83/6
CMS11μj ⊕ ATL11lj	173.02 ± 1.46	0.32	0.36	0.74	0.23	0.00	0.00	0.07	0.76	0.54	0.06	0.24	0.49	0.10	0.10	0.12	0.31	0.64/1
NO CORRELATIONS	173.25 ± 1.05	0.39	0.22	0.48	0.39	0.09	0.02	0.08	0.47	0.29	0.07	0.19	0.32	0.07	0.16	0.11	0.18	1.09/6

- **Largest combined error (most "conservative" result) is found by minimizing information when rescaling correlations independently for the 21 pairs**
  - Result is identical (but for the χ^2) with the combination of CMS11μj and ATL11jl alone
  - Checked that 20 information derivatives are 0 (for all covariances but that of CMS11μj and ATL11jl)
  - Two BLUE coefficients for CMS11μj and ATL11jl are > 0; five others are 0 (remember $dI/dM_{ij} \sim -\lambda_i\lambda_j$)
  - *The most conservative option consists in not using the results with λ<0!*
- **Some negative BLUE coefficients remain for the other two minimizations and (fewer) for the onionization procedure**
- ***All of these procedures change the sys/stat balance, not only the total error!***
  - Nominal correlations: lower total error and systematics, higher statistical error
  - Modified correlations: higher total error and systematics, lower statistical error

PRELIMINARY!
WORK IN PROGRESS!

Recommendations?

CERN IT Department

- **Assume you have chosen your best estimate of correlations ("nominal")**
- **If all λ>0 at "nominal" correlations, you are in a low-correlation regime**
  - Even if correlations are high, their estimates are "conservative" enough: they do not increase the available information, they decrease it – nothing to worry about! ☺
- **If some λ are negative, you are in a high-correlation regime**
  - You should be aware that these high correlations are adding information!
    - Effectively, the measurements with λ <0 are only adding information through these correlations
  - *We advise that all correlations should then be reviewed on a case by case basis*
    - There are legitimate cases where high correlations exist and may help you constrain common systematics (e.g. you varied same MC parameter or depend on a common E_{beam} measurement)
    - But if you *"conservatively"* estimated correlations to be 100%, then you are definitely wrong!
      - If correlations are overestimated, you are certainly underestimating the combined systematics, even more than you are underestimating the total combined error – and your central values are shifted, too
- **What can you do to provide a more realistic (or "conservative") estimate?**
  - Prioritize the correlations to analyze, using information derivatives
    - Then *measure* these correlations with your data and MC if this is possible
  - For correlations you cannot measure precisely, consider some of the tools we presented
    - You may try to minimize information in those correlations alone, or try the onionization procedure (if σ_S for A and B is correlated, and $\sigma_{S,B} > \sigma_{S,A}$, strip out $\sigma^2_{S',B} = \sigma^2_{S,B} - \sigma^2_{S,A}$ as uncorrelated)
  - If you have no clue, omitting the measurements with λ <0 is the most conservative option

Summary:
"relative importance" revisited and conclusions

LHC m_{top} marginal information weights

CERN IT Department

	m_{top} [GeV]	ΔI [1/GeV^2]	λ	$\frac{\lvert\lambda\rvert}{\sum\lvert\lambda\rvert}$	$\frac{\Delta I}{I}$
CMS11μj	172.64 ± 1.53	0.425	84.3%	65.7%	83.8%
ATL11lj	174.53 ± 2.39	0.175	29.9%	23.3%	34.5%
ATL10lj	169.33 ± 6.32	0.025	-6.8%	5.3%	4.9%
CMS11ll	173.30 ± 2.96	0.114	-4.8%	3.7%	22.5%
CMS10ll	175.50 ± 6.49	0.024	-1.9%	1.5%	4.7%
ATL11aj	174.90 ± 4.44	0.051	-0.4%	0.3%	10.0%
CMS10lj	173.10 ± 3.41	0.086	-0.2%	0.2%	16.9%
Correlations	—	-0.392	—	—	-77.4%
Total	173.33 ± 1.40	0.507	100.0%	100.0%	100.0%

Split #0

$$I_{corr} = I - I_A - I_B = \frac{1}{\sigma^2_{\hat{Y}}} - \frac{1}{\sigma^2_A} - \frac{1}{\sigma^2_B}$$

Split low and high correlations by rescaling c as

$$c' = \frac{1}{(\tilde{u}\mathcal{D}^{-1}c)} c$$

For adding ATL10lj, this factor is 0.543

I_{tot}=0.507

Split #1

	m_{top} [GeV]	ΔI [1/GeV^2]	λ	$\frac{\lvert\lambda\rvert}{\sum\lvert\lambda\rvert}$	$\frac{\Delta I}{I}$	$\sum\frac{\Delta I}{I}$
CMS11μj	172.64 ± 1.53	0.425	84.3%	65.7%	83.8%	83.8%
Low corr.	—	—	—	—	—	
ATL11lj	174.53 ± 2.39	0.175	29.9%	23.3%	34.5%	8.6%
Low corr.	—	-0.131	—	—	-25.9%	
ATL10lj	169.33 ± 6.32	0.025	-6.8%	5.3%	4.9%	5.8%
Low corr.	—	-0.025	—	—	-4.9%	
High corr.	—	0.030	—	—	5.8%	
CMS11ll	173.30 ± 2.96	0.114	-4.8%	3.7%	22.5%	1.1%
Low corr.	—	-0.114	—	—	-22.5%	
High corr.	—	0.006	—	—	1.1%	
CMS10ll	175.50 ± 6.49	0.024	-1.9%	1.5%	4.7%	0.7%
Low corr.	—	-0.024	—	—	-4.7%	
High corr.	—	0.003	—	—	0.7%	
ATL11aj	174.90 ± 4.44	0.051	-0.4%	0.3%	10.0%	0.0%
Low corr.	—	-0.051	—	—	-10.0%	
High corr.	—	0.000	—	—	0.0%	
CMS10lj	173.10 ± 3.41	0.086	-0.2%	0.2%	16.9%	0.0%
Low corr.	—	-0.086	—	—	-16.9%	
High corr.	—	0.000	—	—	0.0%	
Total	173.33 ± 1.40	0.507	100.0%	100.0%	100.0%	100.0%

“Split #0”: information weights (IW) from Part I of this talk

- One IW per measurement and one for all correlations together

“Split #1”: one possible way to split up IW for correlations

- Split it up completely amongst all measurements
- Use “marginal information inflow”
  - This requires a **ranking** by BLUE coefficient
  - Two measurements contribute 0.0% information if they are added last

More complex ways to split this up may be more satisfactory

- It would be nice to treat differently negative/positive information contributions in low/high correlation regimes

PRELIMINARY!
WORK IN PROGRESS!

Conclusions

CERN IT Department

- **We propose to use information contributions (instead of absolute BLUE coefficients) to assess the "relative importance" of measurements**
  - One contribution to information comes from the correlations between the measurements and cannot generally be attributed to any one of them individually
  - Marginal information inflow is a possible alternative but implies a ranking

- **Negative BLUE coefficients indicate a "high correlation" regime**
  - It is crucial to properly assess correlations in this regime: overestimated correlations may lead to largely underestimated combined systematics

- **We propose a few tools to help a critical review of correlation estimates**
  - Derivatives of information may help to prioritize the most sensitive correlations
  - Information minimization and covariance onionization may help be more conservative
  - Understanding the marginal inflow of information from the addition of one measurement to the combination may help decide what to do with it
  - **Of course the ideal solution is to measure correlations in the data and MC**

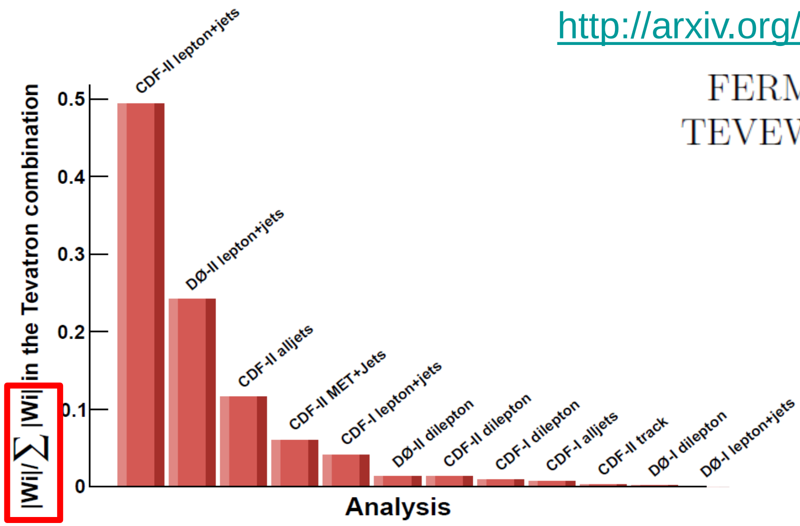
High correlations have a large effect on the combination:
as much effort should go into properly estimating correlations than individual errors

Some useful references

- **A. C. Aitken, *On Least Squares and Linear Combinations of Observations*, Proc. Roy. Soc. Edinburgh 55 (1935), 42**
  - The first published description of the BLUE technique (AFAIK)
- **L. Lyons, D. Gibaut, P. Clifford, *How to combine correlated estimates of a single physical quantity*, Nucl. Instr. Meth. A270 (1988) 110**
  - Extensive discussion of correlations and negative weights for two measurements of a single observable! A very useful read! ☺
- **A. Valassi, *Combining correlated measurements of several different physical quantities*, Nucl. Instr. Meth. A500 (2003) 391**
  - Generalization of Lyons formulas and computation of individual error contributions for many observables (e.g. LEPEWWG)
- **F. James, *Statistical Methods in Experimental Physics (2^{nd} Edition)*, World Scientific (2006)**
  - A very useful textbook covering information and estimation
- **A. van den Bos, *Parameter Estimation for Scientists and Engineers*, Wiley-Interscience (2007)**
  - Another very useful textbook covering information and estimation

RESERVE SLIDES

Negative weights (and absolute weights) @Tevatron



http://arxiv.org/pdf/1107.5255.pdf

FERMILAB-TM-2504-E
TEVEWWG/top 2011/xx
CDF Note 10549
DØ Note 6222

July 2011

	Run I published					Run II published					Run II preliminary	
	CDF			DØ		CDF			DØ		CDF	
	l+jt	di-l	allh	l+jt	di-l	l+jt	di-l	Lxy	l+jt	di-l	allh	Met
Pull	+0.40	−0.51	+1.12	+1.33	−0.37	−0.23	−0.81	−0.67	1.52	0.27	+0.40	−0.36
Weight [%]	−4.7	−1.0	−0.8	−0.0	−0.2	+56.6	+1.4	+0.3	+27.2	+1.5	+14.0	+6.7

The weights of some of the measurements are negative. In general, this situation can occur if the correlation between two measurements is larger than the ratio of their total uncertainties. This is indeed the case here. In these instances the less precise measurement will usually acquire a negative weight. While a weight of zero means that a particular input is effectively ignored in the combination, a negative weight means that it affects the resulting M_t central value and helps reduce the total uncertainty. To visualize the weight each measurement carries in the combination, Fig. 2 shows the absolute values of the weight of each measurement divided by the sum of the absolute values of the weights of all input measurements.

Information

- **General definition for many observables**

$$\mathcal{I}^{(\mathbf{X})}_{\alpha\beta} = E\left[\frac{\partial \log p(\mathbf{y};\mathbf{X})}{\partial X_\alpha}\frac{\partial \log p(\mathbf{y};\mathbf{X})}{\partial X_\beta}\right]$$
$$= \int \frac{\partial \log p(\mathbf{y};\mathbf{X})}{\partial X_\alpha}\frac{\partial \log p(\mathbf{y};\mathbf{X})}{\partial X_\beta}\, p(\mathbf{y};\mathbf{X})\, dy_1 \ldots dy_n$$

- **For multivariate Gaussians (as in BLUE assumption) everything becomes much easier**

$$p(\mathbf{y};\mathbf{X}) = \frac{1}{(2\pi)^{n/2}(\det \mathcal{M})^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(y-\mathcal{U}X)_i \mathcal{M}^{-1}_{ij}(y-\mathcal{U}X)_j\right)$$

- **Information becomes** $$\mathcal{I}^{(\mathbf{X})}_{\alpha\beta} = (\tilde{\mathcal{U}}\mathcal{M}^{-1}\mathcal{U})_{\alpha\beta}$$

related to the BLUE covariance $$\mathrm{cov}(\hat{x}_\alpha\,,\,\hat{x}_\beta) = \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_{\alpha i}\,\mathcal{M}_{ij}\,\lambda_{\beta j} = (\tilde{\mathcal{U}}\mathcal{M}^{-1}\mathcal{U})^{-1}_{\alpha\beta}$$

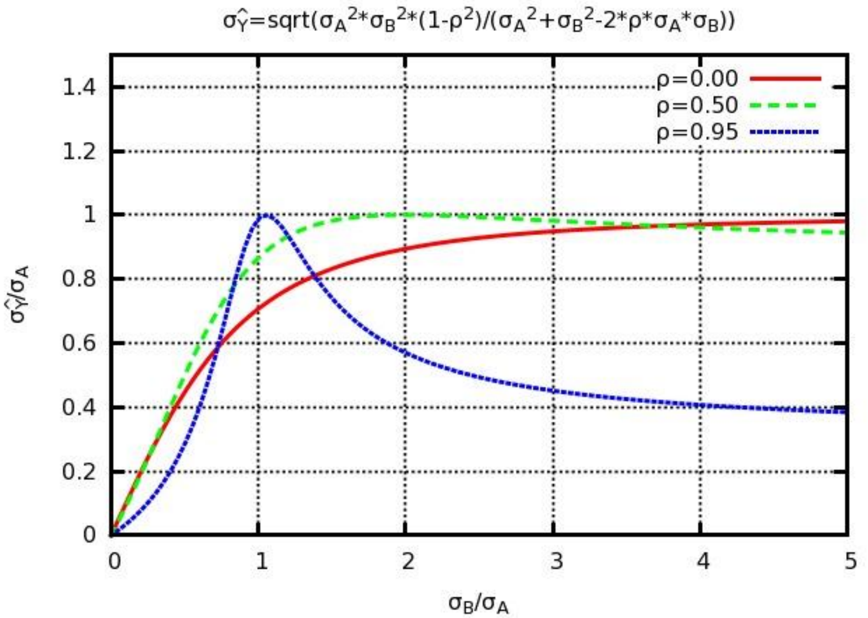
by $$\mathrm{cov}(\hat{x}_\alpha, \hat{x}_\beta) = (\mathcal{I}^{(\mathbf{X})})^{-1}_{\alpha\beta}$$

i.e. BLUE is efficient (satisfies the Cramer Rao lower bound)

What if correlations are fixed a priori?

CERN IT Department

- **Fixing σ_A and ρ and varying σ_B, the result is somewhat surprising, but completely in line with the previous comments:**



- **Once in the high-correlation regime ($\sigma_B > \sigma_A/\rho$), the higher is σ_B, the higher the total information and the lower the combined error!**
  - The better you measure B (as long as $\sigma_B > \sigma_A/\rho$), the worse the combination!
  - The derivative of total information with respect to $(1/\sigma_B)^2$ is negative – it is actually λ_B!

Integrals

- **We also considered using integrals of the dI/df information derivatives to split up the information weight of correlations**
  - The differential of I for a step in the multi dimensional space of rescaling factors can be easily written as the sum of contributions from each off-diagonal element in each error source, individually
  - Integrate this by defining a "path" to transform 0 correlations (all scale factors = 0) to nominal correlations (all scale factors = 1)
  - Use two segments of straight lines, from 0 to a minimum (e.g. minimum by off-diagonal element) and then up to nominal values, to further split up two separate contributions in the low-correlation and high-correlation regimes

- **Some interesting results but in the end why bother?**
  - Derivatives alone provide useful tool to prioritize correlations to re-analyze
  - Marginal information analysis seems to be a more natural and better fit to split up "contributions to knowledge" if really needed
    - series converges quite fast in high-correlation regime because we start from minimum (only $\lambda>0$) and we then add measurements in a carefully chosen order

λ=0 – information vs. χ^2

- **A measurement with a zero weight λ_i=0 in a BLUE combination does not contribute any *information*…**
  – i.e. it does not reduce the variance of the combined estimate

- **… but note that it does contribute to the *χ^2 calculation***
  – the χ^2 is essentially a weighted sum of the products of the differences to the estimate, with weights $(M^{-1})_{ij}$ given by the inverse of the covariance matrix
  – the three sets of "weights", BLUE coefficients, information weights and χ^2 weights are all equal to $1/\sigma_i^2$ if there are no correlations

- **Useful to assess if correlations are correctly estimated**
  – similarly, measurements with λ_i<0 also contribute to the χ^2

The problem with absolute values

CERN IT Department

- Take A, B, C, with *C uncorrelated to A or B*:

$$\begin{pmatrix} \sigma_A^2 & \rho_{AB}\sigma_A\sigma_B & 0 \\ \rho_{AB}\sigma_A\sigma_B & \sigma_B^2 & 0 \\ 0 & 0 & \sigma_C^2 \end{pmatrix}$$

- Combining A, B, C gives the same result as first combining A,B and then adding C:

$$\frac{1}{(\sigma_{\hat{Y}}^2)_{ABC}} = \frac{1}{(\sigma_{\hat{Y}}^2)_{AB}} + \frac{1}{\sigma_C^2}$$

$$(\lambda_A)_{ABC} = \frac{(\lambda_A)_{AB}/(\sigma_{\hat{Y}}^2)_{AB}}{1/(\sigma_{\hat{Y}}^2)_{AB} + 1/\sigma_C^2}$$

$$(\lambda_B)_{ABC} = \frac{(\lambda_B)_{AB}/(\sigma_{\hat{Y}}^2)_{AB}}{1/(\sigma_{\hat{Y}}^2)_{AB} + 1/\sigma_C^2}$$

$$(\lambda_C)_{ABC} = \frac{1/\sigma_C^2}{1/(\sigma_{\hat{Y}}^2)_{AB} + 1/\sigma_C^2}$$

- Intuitively, the relative contribution of C is its "information weight"

$$(\mathrm{IW}_C)_{A,B,C} = \frac{I_C}{I_C + I_{AB}} = \frac{1/\sigma_C^2}{1/(\sigma_{\hat{Y}}^2)_{AB} + 1/\sigma_C^2} = (\lambda_C)_{ABC} = \frac{(\lambda_C)_{ABC}}{(\lambda_A)_{ABC} + (\lambda_B)_{ABC} + (\lambda_C)_{ABC}}$$

=1

while the "relative importance" based on absolute values is

$$(\mathrm{RI}_C)_{A,B,C} = \frac{|(\lambda_C)_{ABC}|}{|(\lambda_A)_{ABC}| + |(\lambda_B)_{ABC}| + |(\lambda_C)_{ABC}|}$$

>1

- Two issues if $\rho_{AB}>\sigma_A/\sigma_B$: first, **RI underestimates the contribution of C**; second, **RI gives different results for (A,B,C) and for (AB,C)**!

$$(\mathrm{RI}_C)_{A,B,C} < (\mathrm{RI}_C)_{AB,C} = (\mathrm{IW}_C)_{AB,C} = (\mathrm{IW}_C)_{A,B,C} \quad \text{if } \rho > \sigma_A/\sigma_B$$

Two measurements of 1 observable

- **For two measurements A, B with errors $\sigma_A < \sigma_B$ and correlation ρ:**

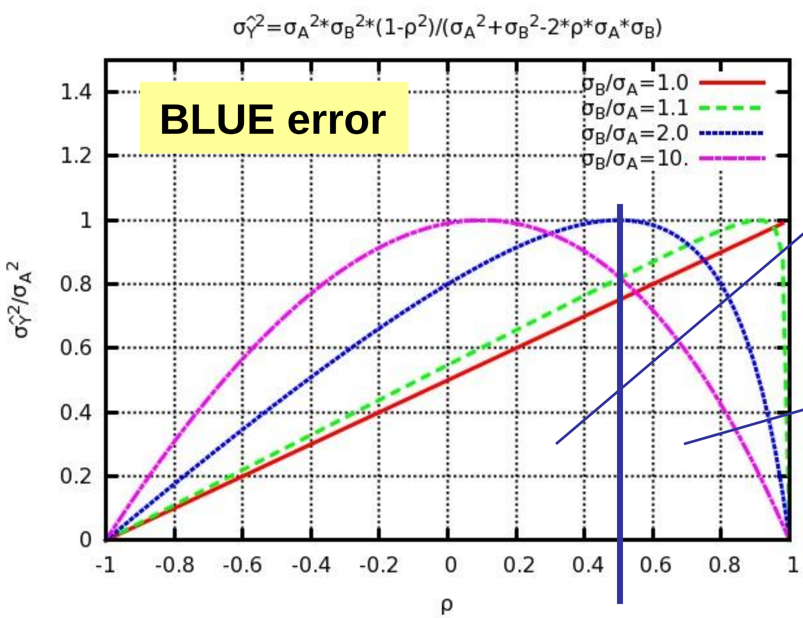
$$\hat{Y} = \lambda_A y_A + \lambda_B y_B$$

$$\sigma_{\hat{Y}}^2 = \frac{\sigma_A^2 \sigma_B^2 (1-\rho^2)}{\sigma_A^2 + \sigma_B^2 - 2\rho\sigma_A\sigma_B} = \frac{1}{I}$$

$$\lambda_A = \frac{\sigma_B^2 - \rho\sigma_A\sigma_B}{\sigma_A^2 + \sigma_B^2 - 2\rho\sigma_A\sigma_B}$$

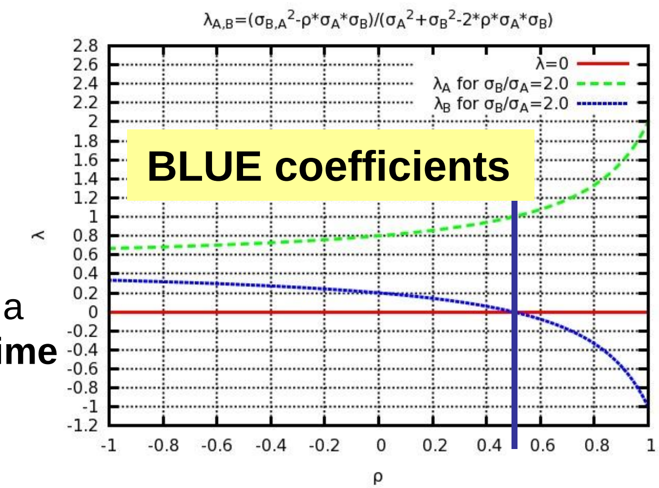
$$\lambda_B = \frac{\sigma_A^2 - \rho\sigma_A\sigma_B}{\sigma_A^2 + \sigma_B^2 - 2\rho\sigma_A\sigma_B}$$

- **The effect of correlations was extensively discussed by *Lyons et al*.**
  - *Fixing σ_A and σ_B and varying ρ, the combined error has a maximum (equal to σ_A) for $\rho=\sigma_A/\sigma_B$, which is where λ_B flips sign from >0 to <0*



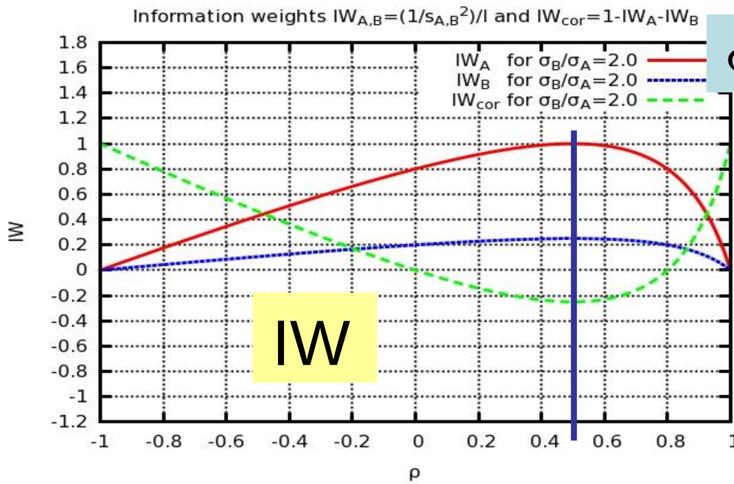
For σ_B/σ_A=2, ρ<0.5 is a **low-correlation regime** where all λ>0

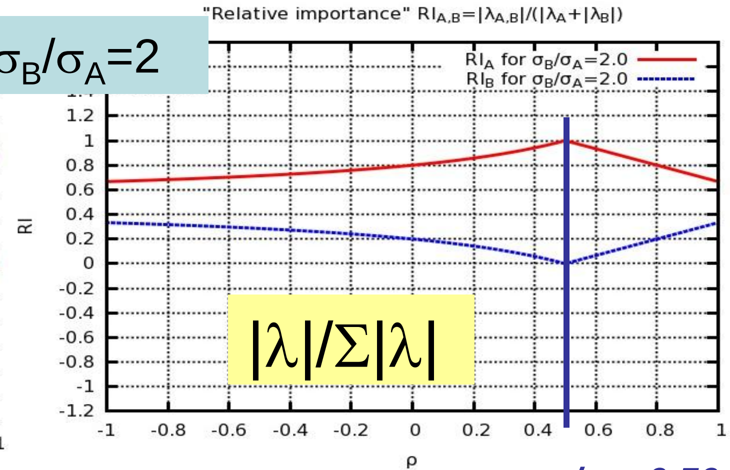
For σ_B/σ_A=2, ρ>0.5 is a **high-correlation regime** where λ_B<0



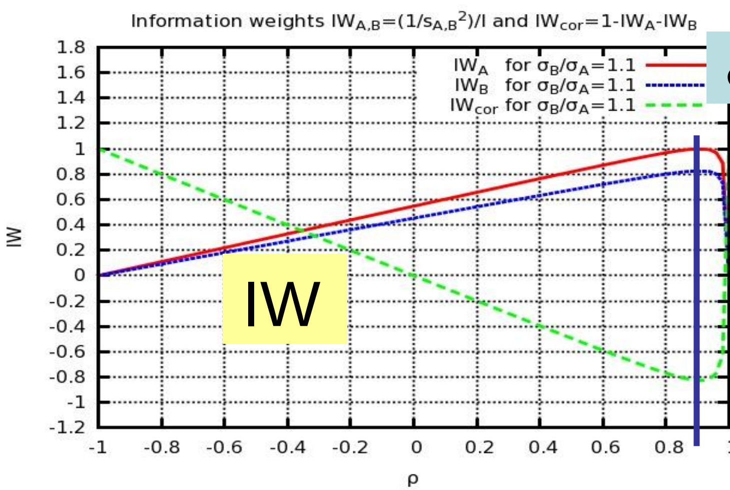
IW's in two measurements of 1 observable

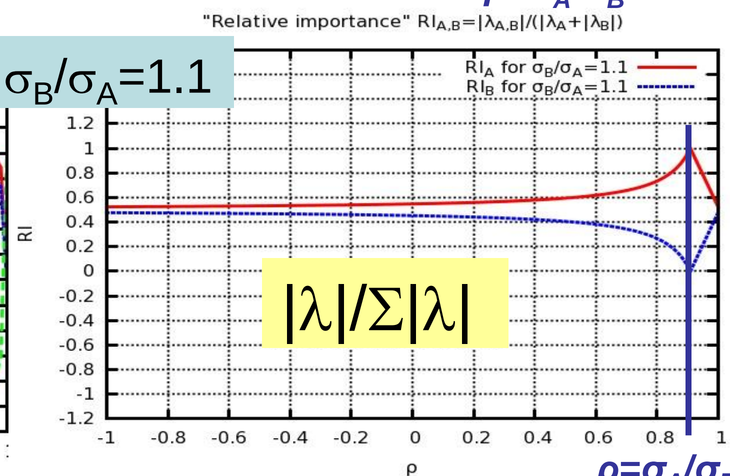
By construction the sum of all information weights (N meas. + 1 correlation) is = 1





$\rho=\sigma_A/\sigma_B=0.50$





$\rho=\sigma_A/\sigma_B=0.91$

Most "conservative" correlation: largest combined error, minimum (largest negative) information weight from correlations

Two measurements – interpretation

CERN IT Department

- **Low correlation regime, $\lambda_B>0$ (i.e. $\sigma_A^2 - \rho\sigma_A\sigma_B > 0$)** >0 (positive definite matrix)
  - The covariance can be seen as the sum of a common error and an uncorrelated error

$$\begin{pmatrix} \sigma_A^2 & \rho\sigma_A\sigma_B \\ \rho\sigma_A\sigma_B & \sigma_B^2 \end{pmatrix} = \begin{pmatrix} \rho\sigma_A\sigma_B & \rho\sigma_A\sigma_B \\ \rho\sigma_A\sigma_B & \rho\sigma_A\sigma_B \end{pmatrix}_{\mathrm{com}} + \begin{pmatrix} \sigma_A^2 - \rho\sigma_A\sigma_B & 0 \\ 0 & \sigma_B^2 - \rho\sigma_A\sigma_B \end{pmatrix}_{\mathrm{unc}}$$

  - You can combine based on the uncorrelated error (compute statistical weights from basic error propagation) and add the common error only at the end
    - *Remember: this happens if the off-diagonal covariance $\rho\sigma_A\sigma_B$ is smaller than the smaller of the two variances σ_A^2 (this will be used later in our "onionization" prescription proposal)*
  - **Adding the less precise measurement B to the combination brings in additional information because B contributes independent (uncorrelated) knowledge about the unknown parameter**

- **High correlation regime, $\lambda_B<0$ (i.e. $\sigma_A^2 - \rho\sigma_A\sigma_B < 0$)**
  - *The covariance can NOT be seen as the sum of a common error and an uncorrelated error*
    - Subtracting a common error matrix $\rho\sigma_A\sigma_B$ from the covariance would give a non-positive-definite matrix
    - The covariance can instead be seen as the sum of a common error and of a 100% positively correlated effect
  - **Adding the less precise measurement B to the combination brings in additional information because B helps to constrain a correlated error on A to which it has a different sensitivity**
    - There are cases where this is perfectly legitimate: *you may leverage on the different sensitivity of two measurements to a common background or a common MC parameter to reduce the uncertainty on them*
    - But this may effectively (and incorrectly) happen *also if correlations are being overestimated*

- **Boundary is $\lambda_B=0$ – where B brings no additional information**
  - **Adding the less precise measurement B to the combination brings in NO additional information because its error has a part common with A plus an additional "uncorrelated" component**

Two measurements – interpretation

CERN IT Department

- **Low correlation regime, $\lambda_B>0$ (i.e. $\sigma_A^2 - \rho\sigma_A\sigma_B > 0$)** ≥0 (positive definite matrix)
  - The covariance can be seen as the sum of a common error and an uncorrelated error

$$\begin{pmatrix} \sigma_A^2 & \rho\sigma_A\sigma_B \\ \rho\sigma_A\sigma_B & \sigma_B^2 \end{pmatrix} = \begin{pmatrix} \rho\sigma_A\sigma_B & \rho\sigma_A\sigma_B \\ \rho\sigma_A\sigma_B & \rho\sigma_A\sigma_B \end{pmatrix}_{\mathrm{com}} + \begin{pmatrix} \sigma_A^2 - \rho\sigma_A\sigma_B & 0 \\ 0 & \sigma_B^2 - \rho\sigma_A\sigma_B \end{pmatrix}_{\mathrm{unc}}$$

  - You can combine based on the uncorrelated error (compute statistical weights from basic error propagation) and add the common error only at the end
    - *Remember: this happens if the off-diagonal covariance $\rho\sigma_A\sigma_B$ is smaller than the smaller of the two variances σ_A^2 (this will be used later in our "onionization" prescription proposal)*
  - **Adding the less precise measurement B to the combination brings in additional information because B contributes independent (uncorrelated) knowledge about the unknown parameter**

$$\begin{aligned} y_A &= Y_{\mathrm{true}} + \delta_{\mathrm{com}} - \lambda_B \delta_{\mathrm{cor}} \\ y_B &= Y_{\mathrm{true}} + \delta_{\mathrm{com}} + \lambda_A \delta_{\mathrm{cor}} \end{aligned}$$

- **High correlation regime, $\lambda_B<0$ (i.e. $\sigma_A^2 - \rho\sigma_A\sigma_B < 0$)**
  - *The covariance can NOT be seen as the sum of a common error and an uncorrelated error*
    - Subtracting a common error matrix $\rho\sigma_A\sigma_B$ from the covariance would give a non-positive-definite matrix
  - The covariance can be seen as the sum of a common error and of a 100% positively correlated effect

$$\begin{pmatrix} \sigma_A^2 & \rho\sigma_A\sigma_B \\ \rho\sigma_A\sigma_B & \sigma_B^2 \end{pmatrix} = \begin{pmatrix} \sigma_{\hat{Y}}^2 & \sigma_{\hat{Y}}^2 \\ \sigma_{\hat{Y}}^2 & \sigma_{\hat{Y}}^2 \end{pmatrix}_{\mathrm{com}} + (\sigma_A^2 + \sigma_B^2 - 2\rho\sigma_A\sigma_B) \begin{pmatrix} \lambda_B^2 & -\lambda_A\lambda_B \\ -\lambda_A\lambda_B & \lambda_A^2 \end{pmatrix}_{\mathrm{cor}}$$

>0

  - **Adding the less precise measurement B to the combination brings in additional information because B helps to constrain a correlated error on A to which it has a different sensitivity**

- **Boundary is $\lambda_B=0$ – where B brings no additional information**
  - **Adding the less precise measurement B to the combination brings in NO additional information because its error has a part common with A plus an additional "uncorrelated" component**

Marginal information inflow

CERN IT Department

- **How much info does the Nth measurement add to a combination with N-1?**
  - Call M the NxN covariance and D the (N-1)x(N-1) covariance
  - Call c the (N-1) dimensional vector of covariances of the Nth measurement
  - Call d the variance of the Nth measurement
  - Define u as an (N-1) dimensional vector of 1's – contraction by this vector is a sum on N-1 indices

$$\mathcal{M} = \begin{pmatrix} \mathcal{D} & \mathbf{c} \\ \tilde{\mathbf{c}} & d \end{pmatrix}$$

- **The marginal information from the Nth measurement is always ≥ 0** ***[van den Bos]***
  - The information inflow ΔI from the Nth measurement is minimum for (i.e. the Nth measurement brings no marginal information)

$$\Delta I = 0 \iff (\tilde{\mathbf{u}}\mathcal{D}^{-1}\mathbf{c}) - 1 = 0$$

  - This is a *hyperplane in the N-1 dimensional space of correlations of the Nth to all other measurements*

- **But the BLUE coefficient for the Nth measurement is** $\lambda_N \propto -\left((\tilde{\mathbf{u}}\mathcal{D}^{-1}\mathbf{c}) - 1\right)$
  - *In other words, $\Delta I=0$ implies $\lambda_N=0$ and viceversa!*

- **The hyperplane $\lambda_N=0$ separates the space of correlations into two half-spaces:**
  - Low correlation regime (including c=0), where $\lambda_N>0$
    - Here ΔI>0 because the Nth measurement adds independent (uncorrelated) information
  - High correlation regime (not including c=0), where $\lambda_N<0$
    - Here ΔI>0 because the high correlations of the Nth measurement help reduce common systematics
    - Note that rescaling all covariances downwards by a common scale factor to would effectively reduce both λ_N and ΔI to 0

$$\mathbf{c}' = \frac{1}{(\tilde{\mathbf{u}}\mathcal{D}^{-1}\mathbf{c})}\mathbf{c}$$

Information inflow

CERN IT Department

- **How much additional information does the n^th measurement add to the previous combination with only n-1?**

- **The inverse of the covariance matrix** $\mathcal{M} = \begin{pmatrix} \mathcal{D} & \mathbf{c} \\ \tilde{\mathbf{c}} & d \end{pmatrix}$

is $\mathcal{M}^{-1} = \begin{pmatrix} \mathcal{D}^{-1} + \frac{(\mathcal{D}^{-1}\mathbf{c})(\tilde{\mathbf{c}}\mathcal{D}^{-1})}{d-(\tilde{\mathbf{c}}\mathcal{D}^{-1}\mathbf{c})} & \frac{(\mathcal{D}^{-1}\mathbf{c})}{d-(\tilde{\mathbf{c}}\mathcal{D}^{-1}\mathbf{c})} \\ \frac{-(\tilde{\mathbf{c}}\mathcal{D}^{-1})}{d-(\tilde{\mathbf{c}}\mathcal{D}^{-1}\mathbf{c})} & \frac{1}{d-(\tilde{\mathbf{c}}\mathcal{D}^{-1}\mathbf{c})} \end{pmatrix}$

- **Defining ũ as a vector of 1's, it can be shown** ***[van den Bos]*** **that the information contributed by the n^th measurement is**

$$\Delta I = \frac{\left((\tilde{\mathbf{u}}\mathcal{D}^{-1}\mathbf{c}) - 1\right)^2}{d - (\tilde{\mathbf{c}}\mathcal{D}^{-1}\mathbf{c})} \geq 0$$

i.e. the inflow of information is always non-negative

  – It is simply equal to 1/d (inverse of variance of n^th measurement) if all correlations are 0

Information inflow and correlations

CERN IT Department

- **The inflow ΔI is zero for** $\Delta I = 0 \iff (\tilde{\mathbf{u}}\mathcal{D}^{-1}\mathbf{c}) - 1 = 0$
  - This defines a hyperplane in the n-1 dimensional space of covariances *c* between the n^{th} and all other measurements
- **But** $\boldsymbol{\lambda} = \frac{\tilde{\mathrm{U}}\mathcal{M}^{-1}}{I}$ **means that** $\lambda_n = -\frac{1}{I} \times \frac{(\tilde{\mathbf{u}}\mathcal{D}^{-1}\mathbf{c}) - 1}{d - (\tilde{\mathbf{c}}\mathcal{D}^{-1}\mathbf{c})}$
  - *In other words, ΔI=0 implies λ_n=0 and viceversa!*
- **The hyperplane λ_n=0 (ΔI=0) separates the space of covariances *c* into two half-spaces:** λ_N
  - Low correlation regime (including c=0), where $\lambda_n>0$
    - Here ΔI>0 because the n^{th} measurement adds independent information
  - High correlation regime (not including c=0), where $\lambda_n<0$

$$\lambda_n \leq 0 \iff (\tilde{\mathbf{u}}\mathcal{D}^{-1}\mathbf{c}) - 1 \geq 0$$

    - Here ΔI>0 because correlations help reduce common systematics!
  - If one suspects that correlations are overestimated, rescaling downwards the covariance by $\mathbf{c}' = \frac{1}{(\tilde{\mathbf{u}}\mathcal{D}^{-1}\mathbf{c})}\mathbf{c}$ effectively reduces both λ and ΔI to 0
- **This generalizes to n measurements the discussion of correlations for 2 measurements in the *Lyons et al.* paper**

Information sensitivity to ρ_{ij}

- **What is the effect on Fisher's information (i.e. on the combined BLUE error) of a change in the correlation ρ_{ij} between two measurements y_i and y_j?**
  - Keep the N variances σ_i fixed and vary only the Nx(N-1)/2 correlations ρ_{ij}
  - Using matrix derivatives *[van den Bos]*, it is easy to show that the derivative of the information with respect to the correlation ρ_{ij} is $$\frac{\partial I}{\partial \rho_{ij}} = -2I^2 \lambda_i \lambda_j \sqrt{\sigma_i}\sqrt{\sigma_j}$$

- **The Nx(N-1)/2 dimensional space of correlations is split into two regimes:**
  - Low correlation regime (including "ρ_{ij}=0 for all ij"), where all $\lambda>0$
    - All derivatives are negative: increasing *any* correlation decreases the information (i.e. increasing any correlation increases the combined error – correlations "conservative")
  - High correlation regime (not including "ρ_{ij}=0 for all ij"), where some $\lambda<0$
    - Some derivatives are positive: increasing those correlations increases the information (i.e. increasing those correlations decreases the combined error – correlations not "conservative")
  - The boundary between the two regimes is a hypersurface where some of the λ are = 0
    - Information is locally minimized (derivatives are 0) with respect to some correlations
    - We checked numerically in some examples (see next slides) that the global minimum of the information (the "most conservative choice of correlations") requires that all λ are either > 0 or = 0: in other words, *a conservative option consists in removing the measurements with $\lambda<0$*

Information sensitivity to M_{ij}

- **Keeping in mind that** $\frac{\partial \mathcal{A}^{-1}}{\partial v_p} = -\mathcal{A}^{-1}\frac{\partial \mathcal{A}}{\partial v_p}\mathcal{A}^{-1}$ ***[van den Bos]*** **and** $I = \tilde{\mathbf{U}}\mathcal{M}^{-1}\mathbf{U}$ **, the partial derivatives of information with respect to a vector *v_p* are:**

$$\frac{\partial I}{\partial v_p} = -\tilde{\mathbf{U}}\mathcal{M}^{-1}\frac{\partial \mathcal{M}}{\partial v_p}\mathcal{M}^{-1}\mathbf{U} = -I^2\tilde{\boldsymbol{\lambda}}\frac{\partial \mathcal{M}}{\partial v_p}\boldsymbol{\lambda}$$

- **Deriving with respect to the *off-diagonal* element $M_{ij}=M_{ji}$ (i≠j) yields**

$$\boxed{\frac{\partial I}{\partial \mathcal{M}_{ij}} = -2I^2\lambda_i\lambda_j}$$ **because** $$\left(\frac{\partial \mathcal{M}}{\partial \mathcal{M}_{ij}}\right)_{kl} = \delta_{il}\delta_{jk} + \delta_{ik}\delta_{jl} = \begin{cases} 1 & \text{if } i=k \text{ and } j=l\,, \\ 1 & \text{if } i=l \text{ and } j=k\,, \\ 0 & \text{otherwise}\,, \end{cases}$$

- **This shows again that the n(n-1)/2 dimensional space of correlations is split into two regimes:**
  - Low correlation regime (including "M_{ij}=0 for all ij"), where all λ>0
    - All derivatives are negative: increasing *any* correlation will decrease information
  - High correlation regime (not including "M_{ij}=0 for all ij"), where some λ<0
    - Some derivatives are positive: increasing those correlations increases information
  - The boundary between the two regimes, where information is minimized as dI/dM=0 ("most conservative" correlations), requires that some λ=0
    - In doubt, a conservative option consists in removing the measurements with λ<0

Is ρ=100% "conservative"?

CERN IT Department

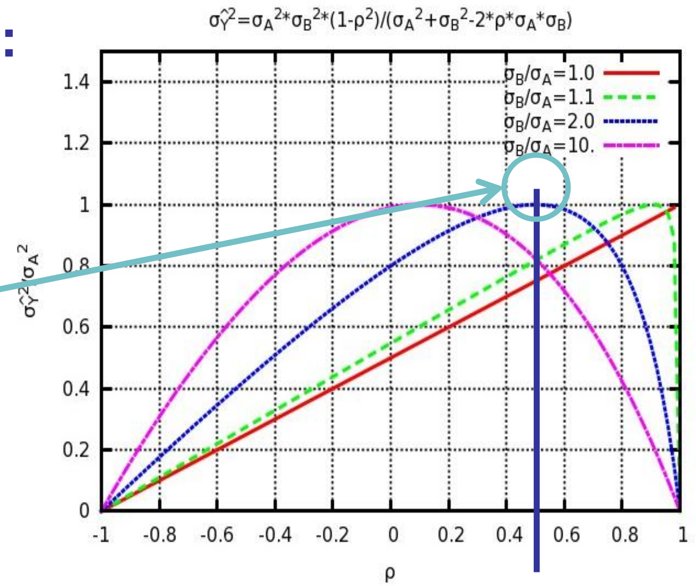
- **Simple case of two measurements. We want:**

$$\rho_{\mathrm{cor}}\sigma_{A,\mathrm{cor}}\sigma_{B,\mathrm{cor}} \leq (\sigma^2_{A,\mathrm{unc}}+\sigma^2_{A,\mathrm{cor}})$$
$$\rho_{\mathrm{cor}}\sigma_{A,\mathrm{cor}}\sigma_{B,\mathrm{cor}} \leq (\sigma^2_{B,\mathrm{unc}}+\sigma^2_{B,\mathrm{cor}})$$

- **Then:**

$$\sigma_{A,\mathrm{cor}} \leq \sigma_{B,\mathrm{cor}}$$

$$\rho_{\mathrm{cor}} \leq \frac{\sigma^2_{A,\mathrm{unc}}+\sigma^2_{A,\mathrm{cor}}}{\sigma_{A,\mathrm{cor}}\sigma_{B,\mathrm{cor}}} = \frac{1}{\left(\frac{\sigma_{A,\mathrm{cor}}}{\sigma_A}\right)^2} \times \frac{\sigma_{A,\mathrm{cor}}}{\sigma_{B,\mathrm{cor}}}$$



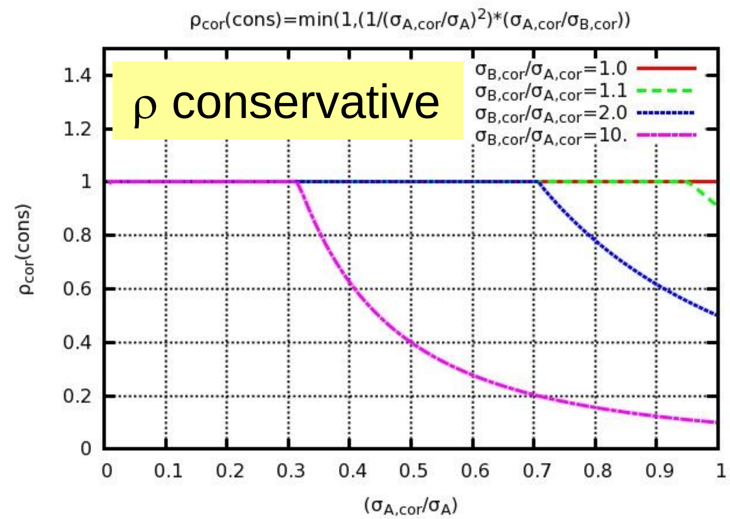
- **Always true for statistically dominated**

$$\sigma_{A,\mathrm{unc}} \gg \sigma_{A,\mathrm{cor}} \text{ and } \sigma_{A,\mathrm{unc}} \gg \sigma_{B,\mathrm{cor}}$$

- **If systematic dominated, rho<1 only if:**

$$(\sigma_{B,\mathrm{cor}}/\sigma_{A,\mathrm{cor}}) \geq (\sigma_A/\sigma_{A,\mathrm{cor}})^2$$

- **For N measurements (rather: unknown correlations) the procedure for finding the maximum error is much less defined**



Derivatives w.r.t correlation re-scale factors

- **Can use derivatives to analyze which correlations (between which 2 measurements) and error sources have largest effect on information**
  - Rescale $\mathcal{M}_{ij}^{[s]} \rightarrow (\mathcal{M}')_{ij}^{[s]} = \begin{cases} f_{ij}^{[s]}\mathcal{M}_{ij}^{[s]} & \text{if } i \neq j, \\ \mathcal{M}_{ii}^{[s]} & \text{if } i = j, \end{cases}$ with $0 \leq f_{ij}^{[s]} \leq 1$
    using a different rescaling factor for each off-diagonal element and error source
  - From $\mathcal{M}'_{ij} = \sum_{s=1}^{S} (\mathcal{M}')_{ij}^{[s]} = \sum_{s=1}^{S} f_{ij}^{[s]}\mathcal{M}_{ij}^{[s]}$ you get $\frac{\partial \mathcal{M}'_{ij}}{\partial f_{ij}^{[s]}} = \mathcal{M}_{ij}^{[s]}$ that implies $\boxed{\frac{\partial I'}{\partial f_{ij}^{[s]}} = -2(I')^2\lambda'_i\lambda'_j\mathcal{M}_{ij}^{[s]}}$

- **In the following we will aggregate these derivatives (21x16 for m_{top}) to reduce complexity, by off-diagonal element (21 for m_{top})** $\frac{\partial I'}{\partial f_{ij}} = \sum_{s}^{S} \frac{\partial I'}{\partial f_{ij}^{[s]}}$ **and error source (16 for m_{top})** $\frac{\partial I'}{\partial f^{[s]}} = \sum_{i<j}^{n} \frac{\partial I'}{\partial f_{ij}^{[s]}}$
  - Total is the derivative by a global scale factor $\frac{\partial I'}{\partial f} = \sum_{s}^{S}\sum_{i<j}^{n} \frac{\partial I'}{\partial f_{ij}^{[s]}} = \sum_{s}^{S} \frac{\partial I'}{\partial f^{[s]}} = \sum_{i<j}^{n} \frac{\partial I'}{\partial f_{ij}}$
  - We will normalize the derivatives to the total information at nominal correlations, so that the derivatives become a-dimensional numbers

Minimizations

- **We studied two procedures to modify the nominal covariance and reduce correlations (e.g. because they have been overestimated)**
  - Rescaling correlations as described before and minimizing information
  - A procedure called "onionization" that we describe later on

- **We minimize information (i.e. maximize the error to be as "conservative" as possible) with respect to the three sets of rescaling factors described in a previous slide**
  - Vary one rescaling factor per error source (16 for m_{top})
    - For a given error source, use the same factor for all off-diagonal elements (i.e. for the whole partial covariance for that error source)
  - Vary one rescaling factor per off-diagonal element (21 for m_{top})
    - For a given off-diagonal element, use the same factor for all error sources
    - This may lead into non-physical space (non positive definite covariances) – give up in that case
  - Vary one global rescaling factor (1 for m_{top})

The "onionization" prescription

- **How can we ensure that** $\sum_{s=1}^{S} \rho_{\mathrm{cor}}^{[s]} \sigma_{\mathrm{A,cor}}^{[s]} \sigma_{\mathrm{B,cor}}^{[s]} \leq (\sigma_{\mathrm{A,unc}}^2 + \sum_{s=1}^{S} (\sigma_{\mathrm{A,cor}}^{[s]})^2)$ **? (*)**

  $\sum_{s=1}^{S} \rho_{\mathrm{cor}}^{[s]} \sigma_{\mathrm{A,cor}}^{[s]} \sigma_{\mathrm{B,cor}}^{[s]} \leq (\sigma_{\mathrm{B,unc}}^2 + \sum_{s=1}^{S} (\sigma_{\mathrm{B,cor}}^{[s]})^2)$

- **We considered:**

$$\begin{aligned} \rho_{\mathrm{cor}}^{[s]} \sigma_{\mathrm{A,cor}}^{[s]} \sigma_{\mathrm{B,cor}}^{[s]} &\leq (\sigma_{\mathrm{A,unc}}^2 + \textstyle\sum_{s'=1}^{S} (\sigma_{\mathrm{A,cor}}^{[s']})^2) \quad \forall s \\ \rho_{\mathrm{cor}}^{[s]} \sigma_{\mathrm{A,cor}}^{[s]} \sigma_{\mathrm{B,cor}}^{[s]} &\leq (\sigma_{\mathrm{B,unc}}^2 + \textstyle\sum_{s'=1}^{S} (\sigma_{\mathrm{B,cor}}^{[s']})^2) \quad \forall s \end{aligned}$$

- **But we prefer the "onionization" of every source of uncertainty: remain ≤**

$$\begin{pmatrix} (\sigma_{A,\mathrm{cor}}^{[s]})^2 & (\sigma_{A,\mathrm{cor}}^{[s]})^2 & (\sigma_{A,\mathrm{cor}}^{[s]})^2 & (\sigma_{A,\mathrm{cor}}^{[s]})^2 & \cdots \\ (\sigma_{A,\mathrm{cor}}^{[s]})^2 & (\sigma_{B,\mathrm{cor}}^{[s]})^2 & (\sigma_{B,\mathrm{cor}}^{[s]})^2 & (\sigma_{B,\mathrm{cor}}^{[s]})^2 & \cdots \\ (\sigma_{A,\mathrm{cor}}^{[s]})^2 & (\sigma_{B,\mathrm{cor}}^{[s]})^2 & (\sigma_{C,\mathrm{cor}}^{[s]})^2 & (\sigma_{C,\mathrm{cor}}^{[s]})^2 & \cdots \\ (\sigma_{A,\mathrm{cor}}^{[s]})^2 & (\sigma_{B,\mathrm{cor}}^{[s]})^2 & (\sigma_{C,\mathrm{cor}}^{[s]})^2 & (\sigma_{D,\mathrm{cor}}^{[s]})^2 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

$$\sigma_{\mathrm{A,cor}} \leq \sigma_{\mathrm{B,cor}} \leq \sigma_{\mathrm{C,cor}} \leq \sigma_{\mathrm{D,cor}}$$

  i.e. :

$$\begin{aligned} \rho_{\mathrm{cor}}^{[s]} \sigma_{A,\mathrm{cor}}^{[s]} \sigma_{B,\mathrm{cor}}^{[s]} &\leq (\sigma_{A,\mathrm{cor}}^{[s]})^2 \quad \forall s \\ \rho_{\mathrm{cor}}^{[s]} \sigma_{A,\mathrm{cor}}^{[s]} \sigma_{B,\mathrm{cor}}^{[s]} &\leq (\sigma_{B,\mathrm{cor}}^{[s]})^2 \quad \forall s \end{aligned}$$

- **However this may transform overestimates of ρ into underestimates**
  - It means *assuming that the difference in sensitivity between measurements to a given systematic effect cannot be used to constrain or determine the actual size of that effect*
    - Which is generally wrong – there are legitimate cases for exploiting this leverage effect
  - And with >2 measurements this condition (*) is not even enough to ensure λ>0 for each λ

- **Of course the only correct way forward is**
  - measure the correlations between measurements (from the data, the MC parameters…)
  - if not possible, assess case by case whether 100% (or e.g. onionizing) is appropriate

Onionization example – RAD @ LHC m_{top}

CMS11ll	0.59	0.61	0.73	0.39	0.91	0.65	0.94
CMS11μj	0.61	0.64	0.75	0.40	0.95	0.68	0.98
CMS10ll	0.73	0.75	0.89	0.48	1.12	0.80	1.15
ATL11lj	0.39	0.40	0.48	1.02	0.60	1.72	2.47
CMS10lj	0.91	0.95	1.12	0.60	1.41	1.01	1.45
ATL11aj	0.65	0.68	0.80	1.72	1.01	2.89	4.16
ATL10lj	0.94	0.98	1.15	2.47	1.45	4.16	5.98

Table 13: Onionization of the RAD covariance. Nominal RAD.

CMS11ll	0.59	0.59	0.59	0.59	0.59	0.59	0.59
CMS11μj	0.59	0.64	0.64	0.64	0.64	0.64	0.64
CMS10ll	0.59	0.64	0.89	0.89	0.89	0.89	0.89
ATL11lj	0.59	0.64	0.89	1.02	1.02	1.02	1.02
CMS10lj	0.59	0.64	0.89	1.02	1.41	1.41	1.41
ATL11aj	0.59	0.64	0.89	1.02	1.41	2.89	2.89
ATL10lj	0.59	0.64	0.89	1.02	1.41	2.89	5.98

Table 14: Onionization of the RAD covariance. Onion upper bound.

CMS11ll	0.59	0.59	0.59	0.39	0.59	0.59	0.59
CMS11μj	0.59	0.64	0.64	0.40	0.64	0.64	0.64
CMS10ll	0.59	0.64	0.89	0.48	0.89	0.80	0.89
ATL11lj	0.39	0.40	0.48	1.02	0.60	1.02	1.02
CMS10lj	0.59	0.64	0.89	0.60	1.41	1.01	1.41
ATL11aj	0.59	0.64	0.80	1.02	1.01	2.89	2.89
ATL10lj	0.59	0.64	0.89	1.02	1.41	2.89	5.98

Table 15: Onionization of the RAD covariance. Onionized RAD.

PRELIMINARY!
WORK IN PROGRESS!

LHC m_{top} – conservative correlations?

CERN IT Department

	Comb. m_{top}	Stat	iJES	bJES	dJES	rJES	Lept	MC	Rad	CR	PDF	DTMO	UE	BGMC	BGDT	Meth	MHI	χ^2/ndof
PUBLISHED	173.33 ± 1.40	0.47	0.38	0.68	0.07	0.06	0.01	0.04	0.69	0.55	0.01	0.19	0.47	0.01	0.16	0.13	0.25	2.49/6
MinimizeGlobalFactor1	173.29 ± 1.41	0.40	0.36	0.71	0.19	0.05	0.01	0.06	0.71	0.53	0.04	0.22	0.47	0.04	0.14	0.12	0.26	2.11/6
MinimizeErrorSrc18	173.22 ± 1.44	0.35	0.35	0.72	0.25	0.02	0.00	0.09	0.73	0.54	0.05	0.24	0.48	0.10	0.12	0.12	0.29	1.97/6
Onionize	173.08 ± 1.44	0.31	0.33	0.72	0.26	0.01	0.00	0.07	0.76	0.54	0.06	0.24	0.47	0.09	0.10	0.11	0.28	1.67/6
MinimizeOffDiagonal21	173.02 ± 1.46	0.32	0.36	0.74	0.23	0.00	0.00	0.07	0.76	0.54	0.06	0.24	0.49	0.10	0.10	0.12	0.31	1.83/6
CMS11μj $\oplus$ ATL11lj	173.02 ± 1.46	0.32	0.36	0.74	0.23	0.00	0.00	0.07	0.76	0.54	0.06	0.24	0.49	0.10	0.10	0.12	0.31	0.64/1
NO CORRELATIONS	173.25 ± 1.05	0.39	0.22	0.48	0.39	0.09	0.02	0.08	0.47	0.29	0.07	0.19	0.32	0.07	0.16	0.11	0.18	1.09/6

Table 9: LHC m_{top} BLUE combinations with nominal and modified correlations. From top to bottom: published results with all 7 measurements and nominal correlations; all correlations rescaled by 1 global factor, maximizing the error; correlations rescaled by 18 scale factors, one per error source, maximizing the error; onionization of the covariance matrix in each error source; correlations rescaled by 21 scale factors, one per off-diagonal element (same factor for all error sources), maximizing the error; results with only the two measurements with positive λ and nominal correlations; all correlations set to 0 everywhere. From top to bottom, the combination essentially moves from a high-correlation regime to a low-correlation regime, possibly passing through the absolute minimum at the minimization by off-diagonal element, which seems to coincide with the combination with two measurements.

- **Absolute minimum is most likely at "MinimizeOffDiagonal21"**
  - Checked that 20 information derivatives are 0 (for all covariances but that of CMS11μj and ATL11jl)
  - Indeed (remember $dI/dM_{ij} \sim -\lambda_i\lambda_j$!) all 5 lambdas become 0 except for CMS11μj and ATL11jl
  - Not surprisingly results coincide with the combination of CMS11μj and ATL11jl alone, but for the χ^2
  - *A very conservative option consists in not using the results with $\lambda<0$!*

- **Some negative BLUE coefficients remain for the other two minimizations and (less) for the onionization procedure**
  - *Note again that all of these procedures change the sys/stat balance, not only the total error!*
  - Some of these procedures may also not be completely physical (covariances not positive definite?)

- ***Carefully measuring correlations is much better than any rule of thumb!***

PRELIMINARY!
WORK IN PROGRESS!

Recommendations?

CERN IT Department

- **Assume you have chosen your best estimate of correlations ("nominal")**
- **If all $\lambda>0$ at "nominal" correlations, you are in a low-correlation regime**
  - Even if correlations are high, their estimates are "conservative" enough: they do not increase the available information, they decrease it – nothing to worry about! ☺
- **If some λ are negative, you are in a high-correlation regime**
  - You should be aware that these high correlations are adding information!
    - Effectively, the measurements with $\lambda<0$ are only adding information through these correlations
  - *We advise that all correlations should then be reviewed on a case by case basis*
    - There are legitimate cases where high correlations exist and may help you constrain common systematics (e.g. you varied same MC parameter or depend on a common E_{beam} measurement)
    - But if you *"conservatively"* estimated correlations to be 100%, then you are definitely wrong!
      - If correlations are overestimated, you are certainly underestimating the combined systematics, even more than you are underestimating the total combined error – and your central values are shifted, too
- **What can you do to provide a more realistic (or "conservative") estimate?**
  - Prioritize the correlations to analyze, using information derivatives
    - Then *measure* these correlations with your data and MC if this is possible
  - For correlations you cannot measure precisely, consider some of the tools we presented
    - You may try to minimize information in those correlations alone, or try the onionization procedure (if σ_S for A and B is correlated, and $\sigma_{S,B} > \sigma_{S,A}$, strip out $\sigma^2_{S',B} = \sigma^2_{S,B}-\sigma^2_{S,A}$ as uncorrelated)
  - If you have no clue, omitting the measurements with $\lambda<0$ is the most conservative option

LHC m_{top} marginal information weights

$$I_{\rm corr} = I - I_{\rm A} - I_{\rm B} = \frac{1}{\sigma^2_{\hat{Y}}} - \frac{1}{\sigma^2_{\rm A}} - \frac{1}{\sigma^2_{\rm B}}$$

	m_{top} [GeV]	ΔI [1/GeV^2]	λ	$\frac{\lvert\lambda\rvert}{\sum\lvert\lambda\rvert}$	$\frac{\Delta I}{I}$	$\sum\frac{\Delta I}{I}$	$\Delta I'$ [1/GeV^2]	$\sum\frac{\Delta I'}{I}$
CMS11μj	172.64 ± 1.53	0.425	84.3%	65.7%	83.8%	83.8%	0.425	74.1%
Low corr.	—	—	—	—	—		-0.049	
ATL11lj	174.53 ± 2.39	0.175	29.9%	23.3%	34.5%	8.6%	0.175	18.4%
Low corr.	—	-0.131	—	—	-25.9%		-0.082	
ATL10lj	169.33 ± 6.32	0.025	-6.8%	5.3%	4.9%	5.8%	0.025	5.8%
Low corr.	—	-0.025	—	—	-4.9%			
High corr.	—	0.030	—	—	5.8%			
CMS11ll	173.30 ± 2.96	0.114	-4.8%	3.7%	22.5%	1.1%	0.006	1.1%
Low corr.	—	-0.114	—	—	-22.5%			
High corr.	—	0.006	—	—	1.1%			
CMS10ll	175.50 ± 6.49	0.024	-1.9%	1.5%	4.7%	0.7%	0.003	0.7%
Low corr.	—	-0.024	—	—	-4.7%			
High corr.	—	0.003	—	—	0.7%			
ATL11aj	174.90 ± 4.44	0.051	-0.4%	0.3%	10.0%	0.0%	0.000	0.0%
Low corr.	—	-0.051	—	—	-10.0%			
High corr.	—	0.000	—	—	0.0%			
CMS10lj	173.10 ± 3.41	0.086	-0.2%	0.2%	16.9%	0.0%	0.000	0.0%
Low corr.	—	-0.086	—	—	-16.9%			
High corr.	—	0.000	—	—	0.0%			
Total	173.33 ± 1.40	0.507	100.0%	100.0%	100.0%	100.0%	0.507	100.0%

Table 7: LHC m_{top} *marginal* information weights. The measurements are listed in the following order: first those with $\lambda > 0$, ordered by decreasing λ; then those with $\lambda < 0$, ordered by decreasing $|\lambda|$. The information inflows ΔI represent the *marginal* contributions of adding to the combination the $n^{\rm th}$ measurement and its correlations to the previous $n-1$ measurements. The "effective" information inflows $\Delta I'$ for the measurements are computed as follows: for those with $\lambda < 0$, they are are equal to ΔI; for those with $\lambda \geq 0$ (the first two only), a BLUE combination of those measurements alone is performed, and the total information contribution from the correlations between them is split up so that the totals for each measurement and its correlations are proportional to the BLUE coefficients (80.1% and 19.9%) in that combination. As a result, the two columns with the information weights computed as $\sum\frac{\Delta I'}{I}$ and $\sum\frac{\Delta I'}{I}$ differ only for $\lambda \geq 0$.

Split low and high correlations by rescaling c as $\mathbf{c}' = \frac{1}{(\tilde{\mathbf{u}}\mathcal{D}^{-1}\mathbf{c})}\mathbf{c}$

For adding ATL10lj, this factor is 0.543

	m_{top} [GeV]	ΔI [1/GeV^2]	λ	$\frac{\lvert\lambda\rvert}{\sum\lvert\lambda\rvert}$	$\frac{\Delta I}{I}$
CMS11μj	172.64 ± 1.53	0.425	84.3%	65.7%	83.8%
ATL11lj	174.53 ± 2.39	0.175	29.9%	23.3%	34.5%
ATL10lj	169.33 ± 6.32	0.025	-6.8%	5.3%	4.9%
CMS11ll	173.30 ± 2.96	0.114	-4.8%	3.7%	22.5%
CMS10ll	175.50 ± 6.49	0.024	-1.9%	1.5%	4.7%
ATL11aj	174.90 ± 4.44	0.051	-0.4%	0.3%	10.0%
CMS10lj	173.10 ± 3.41	0.086	-0.2%	0.2%	16.9%
Correlations	—	-0.392	—	—	-77.4%
Total	173.33 ± 1.40	0.507	100.0%	100.0%	100.0%

These are two possible ways to split up the info contribution from the correlations, based on "marginal" contributions

Note that two measurements contribute 0.0% information

PRELIMINARY!
WORK IN PROGRESS!