# Correlation & Bias in $\gamma$-jet balance

## a talk for the mathematically inclined

Georgios Choudalakis, Eric Feng
University of Chicago

weekly group meeting
2009/05/11

# Notation and Definitions

In derivations,

- $\gamma$ is a short-hand for reconstructed photon $p_T$,
- $j$ is a short-hand for uncorrected jet $p_T$,
- $p_{corr}$ is the jet $p_T$ after the correction. Since we try to adjust the jet to balance the photon, $p_{corr}$ is an estimator for photon's $p_T$, which uses as observable the uncorrected jet $p_T$.
- The notation $<...>_j$ signifies averaging over all events in a given *j*-bin.

**Definition of bias:**

$\hat{\theta}$ is a biased estimator of $\theta$ if $E(\hat{\theta}-\theta) \neq 0$.

**Definition of $k^j_{a,b}$:**

$$\left\langle \frac{a}{b} \right\rangle_j \equiv k^j_{a,b} \frac{\langle a \rangle_j}{\langle b \rangle_j} \Rightarrow k^j_{a,b} = 1 + \mathrm{cov}(a, \frac{1}{b})_j \frac{\langle b \rangle_j}{\langle a \rangle_j}$$

# 4 options of defining & applying the correction that takes us to the $\gamma$ $p_T$

- Option 1: Correction factor: $C = \left\langle \frac{\gamma}{j} \right\rangle_j$

$$p_{\text{corr}} = p_T^{jet}\, C_{\text{jet p}_\text{T}} = j\, C = j \left\langle \frac{\gamma}{j} \right\rangle_j$$

- Option 2:

Correction factor: $C = \left\langle \frac{j}{\gamma} \right\rangle_j$

$$p_{\text{corr}} = j \frac{1}{C} = j \frac{1}{\left\langle \frac{\gamma}{j} \right\rangle_j}$$

- Option 3:

Correction factor: $C = \left\langle \frac{\gamma}{(\gamma + j)/2} \right\rangle_j$

$$p_{\text{corr}} = j \frac{C}{2 - C} = j \frac{\left\langle \frac{\gamma}{j+\gamma} \right\rangle_j}{1 - \left\langle \frac{\gamma}{j+\gamma} \right\rangle_j}$$

- Option 4:

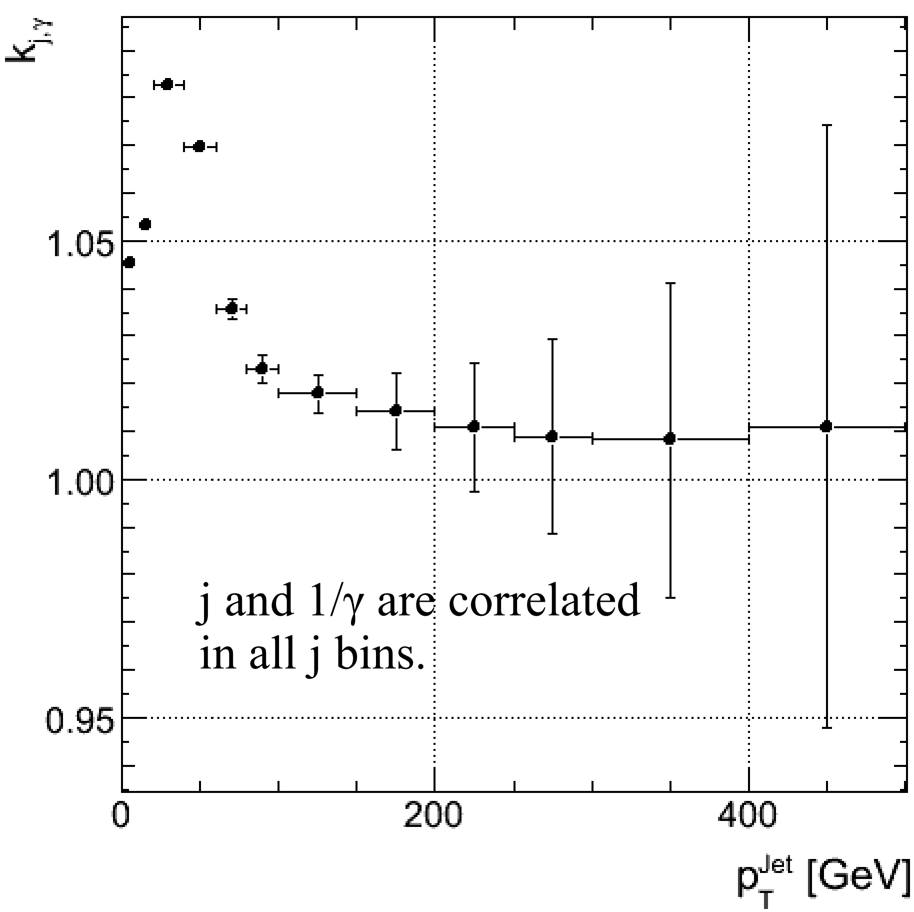
Correction factor: $C = \left\langle \frac{j}{(\gamma + j)/2} \right\rangle_j$

$$p_{\text{corr}} = j \frac{2 - C}{C} = j \frac{1 - \left\langle \frac{j}{j+\gamma} \right\rangle_j}{\left\langle \frac{j}{j+\gamma} \right\rangle_j}$$

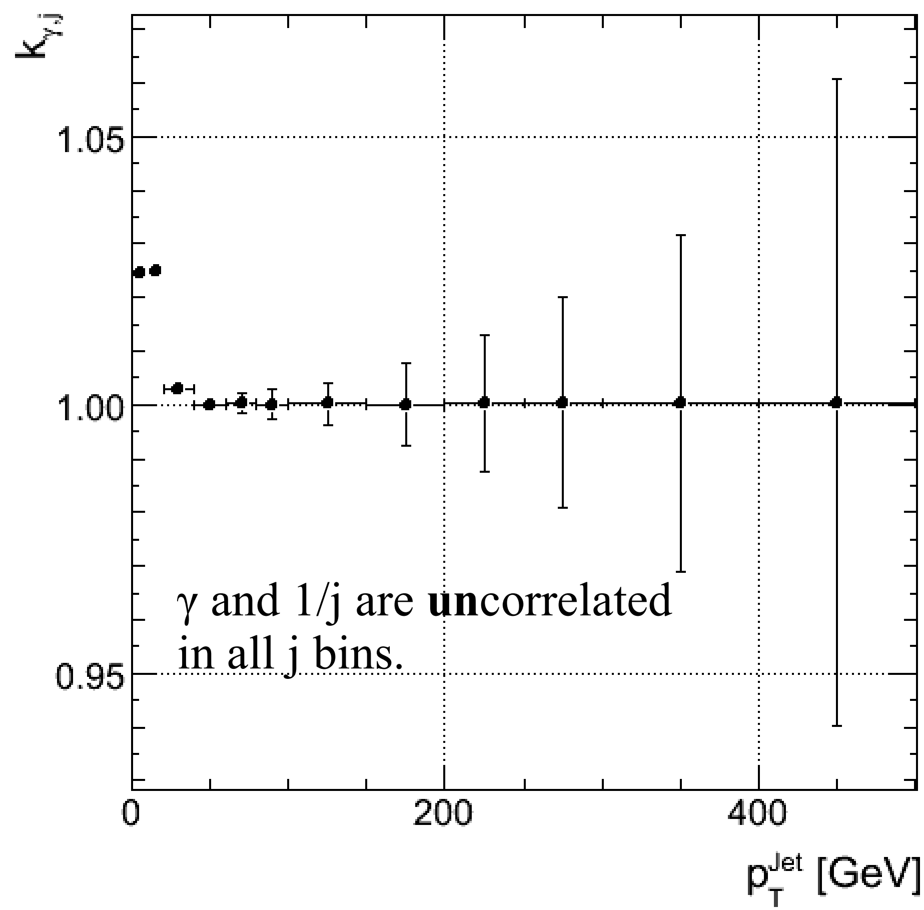
# Correlations

## (using MC, but these are observable in data too)

$k_{j,\gamma} = \langle j/\gamma \rangle_j \,/\, (\langle j \rangle_j / \langle \gamma \rangle_j)$ in each jet uncorrected $p_T$ bin.




$k_{\gamma,j} = \langle \gamma/j \rangle_j \,/\, (\langle \gamma \rangle_j / \langle j \rangle_j)$ in each jet uncorrected $p_T$ bin.

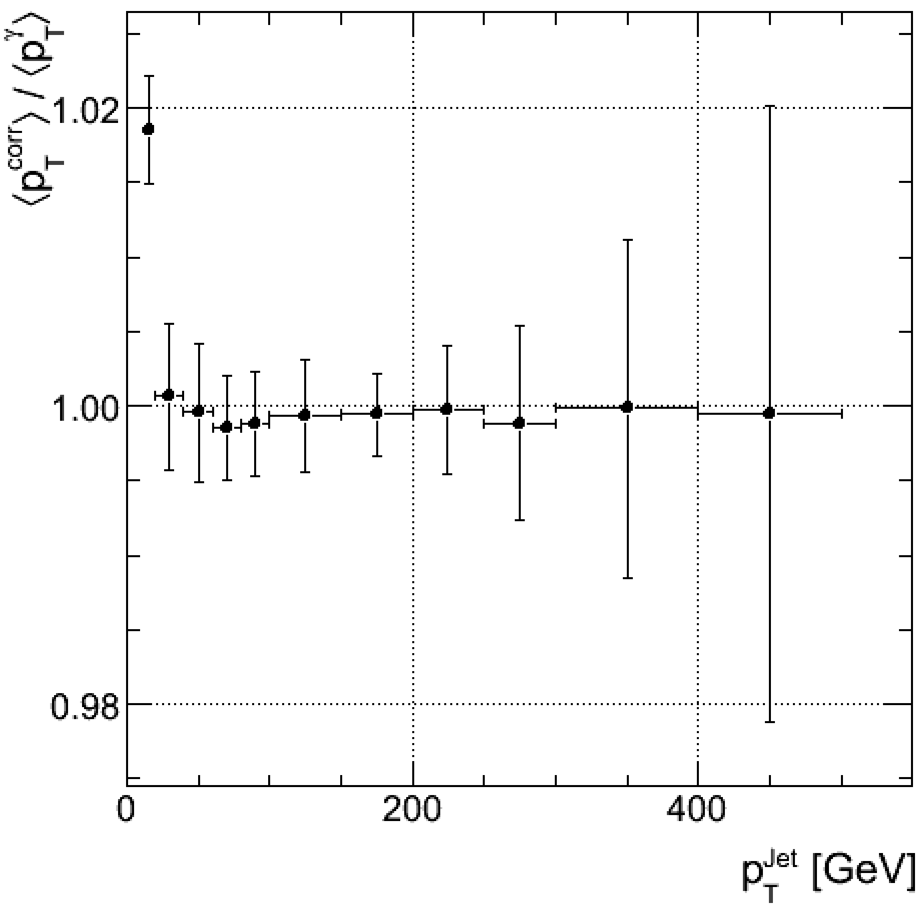


# Option 1

Correction factor: $C = \left\langle \frac{\gamma}{j} \right\rangle_j$

$$\langle p_{\mathrm{corr}} \rangle_j = \langle jC \rangle_j = \left\langle j \left\langle \frac{\gamma}{j} \right\rangle_j \right\rangle_j = \left\langle \frac{\gamma}{j} \right\rangle_j \langle j \rangle_j = k^j_{\gamma,j} \frac{\langle \gamma \rangle_j}{\langle j \rangle_j} \langle j \rangle_j = k^j_{\gamma,j} \langle \gamma \rangle_j$$

If γ and j are uncorrelated, then k=1 and $<p_{corr}>_j = <\gamma>_j$.
In that case $p_{corr}$ is an unbiased estimator of the γ $p_T$, since $<p_{corr} - \gamma> = <\gamma> - <\gamma> = 0$.
But, correlation between γ and j $p_T$ causes the estimator to be biased: $<p_{corr}>/<\gamma> = k_{\gamma,j}$.
$k_{\gamma,j} \approx 1$, so the bias is tiny, even if we do nothing to correct it.




But in general we can correct it by redefining C:

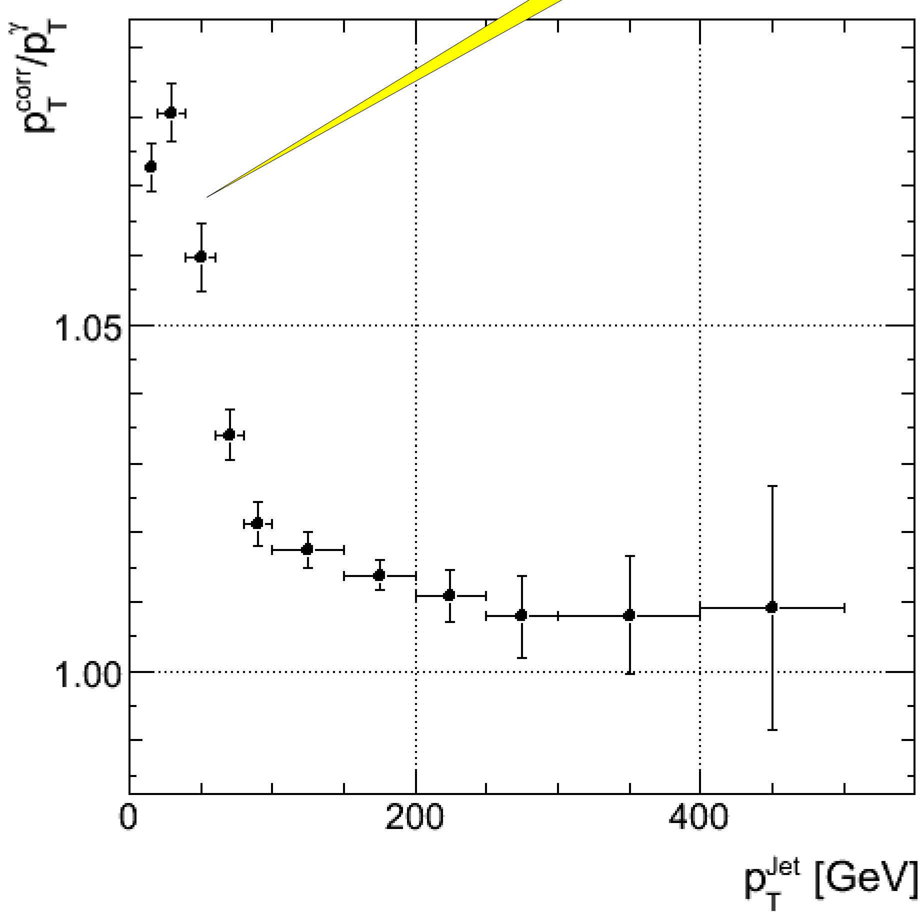
$$C = \left\langle \frac{\gamma}{j} \right\rangle_j \to C' = \frac{C}{k^j_{\gamma,j}} = \frac{\left\langle \frac{\gamma}{j} \right\rangle_j}{\frac{\left\langle \frac{\gamma}{j} \right\rangle_j}{\frac{\langle \gamma \rangle_j}{\langle j \rangle_j}}} = \frac{\langle \gamma \rangle_j}{\langle j \rangle_j}$$

We'll come back to this later

# Option 1 (continued)

$$\left\langle \frac{p_{corr}}{\gamma} \right\rangle_j = \left\langle \frac{jC}{\gamma} \right\rangle_j = \left\langle \frac{j\left\langle \frac{\gamma}{j} \right\rangle_j}{\gamma} \right\rangle_j = \left\langle \frac{\gamma}{j} \right\rangle_j \left\langle \frac{j}{\gamma} \right\rangle_j = k^j_{\gamma,j} \frac{\langle\gamma\rangle_j}{\langle j\rangle_j} k^j_{j,\gamma} \frac{\langle j\rangle_j}{\langle\gamma\rangle_j} = k^j_{\gamma,j} \cdot k^j_{j,\gamma}$$

Before thinking carefully, I thought it would be enough to show $< p_{corr} / \gamma >$. I was hoping that if that were = 1, then I would have shown my estimator is unbiased. WRONG! This is not the same as $< p_{corr} >/< \gamma >$, which is the actual definition of bias.




$< p_{corr}/\gamma >$ will be $\neq 1$, even if $< p_{corr} > = < \gamma >$.

**Since $k_{\gamma,j} \approx 1$, $< p_{corr} / \gamma > \approx 1 \times k_{j,\gamma}$.**

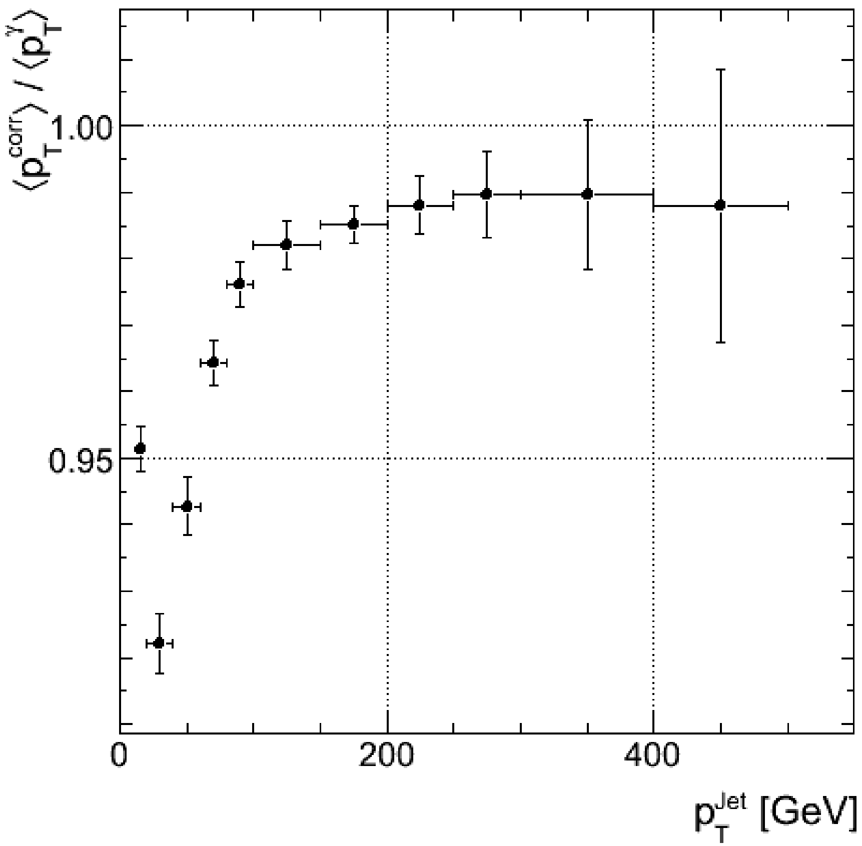
# Option 2

Correction factor: $C = \left\langle \frac{j}{\gamma} \right\rangle_j$

$$\langle p_{\text{corr}} \rangle_j = \langle j/C \rangle_j = \left\langle \frac{j}{\left\langle \frac{j}{\gamma} \right\rangle_j} \right\rangle_j = \frac{\langle j \rangle_j}{\left\langle \frac{j}{\gamma} \right\rangle_j} = \frac{\langle j \rangle_j}{k^j_{j,\gamma} \frac{\langle j \rangle_j}{\langle \gamma \rangle_j}} = \frac{\langle \gamma \rangle_j}{k^j_{j,\gamma}}$$

There is bias.
Unlike Option 1, with Option 2 the bias is large.

**Moral:** Option 2 and Option 1 are not equivalent.
*Correlation beats intuition.*




The bias could be corrected by a redefinition of C:

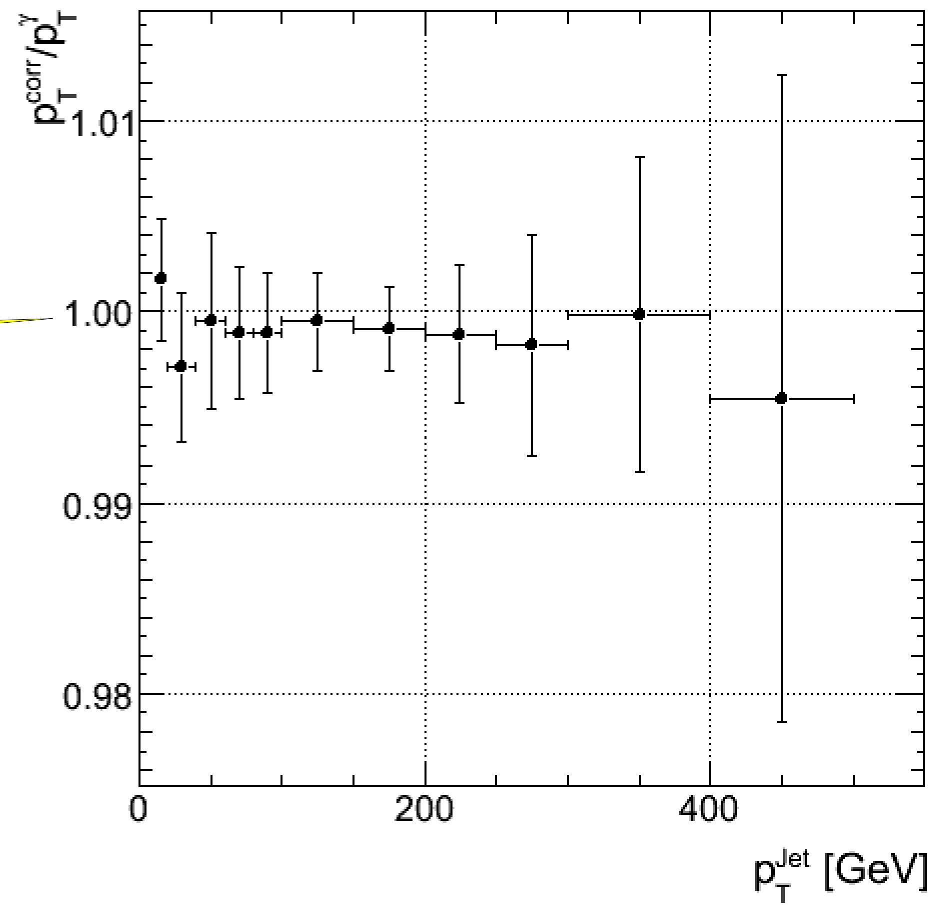
$$C = \left\langle \frac{j}{\gamma} \right\rangle_j \rightarrow C' = \frac{C}{k^j_{j,\gamma}} = \frac{\left\langle \frac{j}{\gamma} \right\rangle_j}{\frac{\left\langle \frac{j}{\gamma} \right\rangle_j}{\frac{\langle j \rangle_j}{\langle \gamma \rangle_j}}} = \frac{\langle j \rangle_j}{\langle \gamma \rangle_j}$$

We'll come back to this later

# Option 2 (continued)

$$\left\langle \frac{p_{\text{corr}}}{\gamma} \right\rangle_j = \left\langle \frac{j/C}{\gamma} \right\rangle_j = \left\langle \frac{j}{\gamma} \frac{1}{\left\langle \frac{j}{\gamma} \right\rangle_j} \right\rangle_j$$

$$= \left\langle \frac{j}{\gamma} \right\rangle_j \frac{1}{\left\langle \frac{j}{\gamma} \right\rangle_j} = 1$$

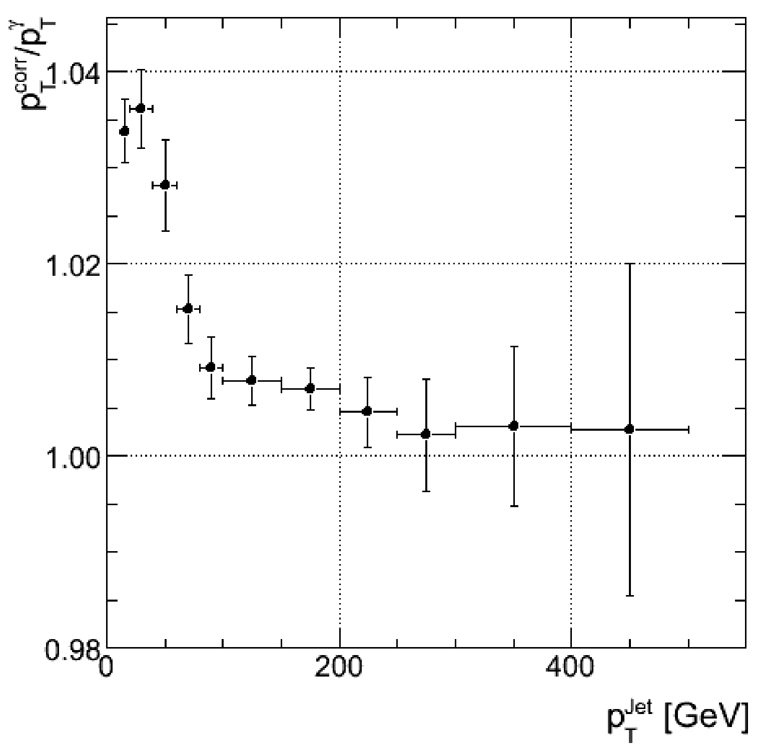


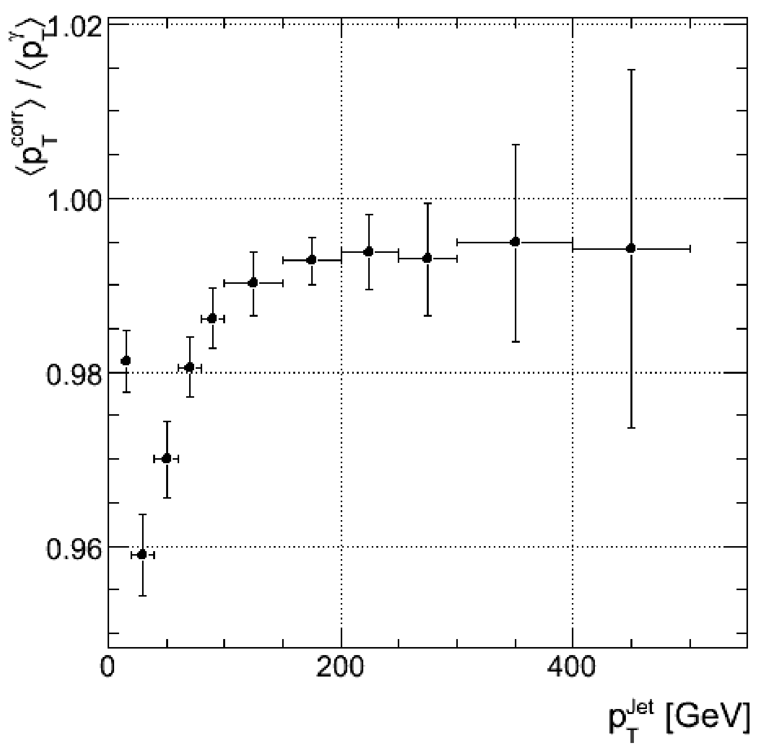

**There is bias** : $\langle p_{corr} \rangle / \langle \gamma \rangle = 1/k_{j,\gamma}$.
and yet in this plot it seems like there isn't.
*This plot is misleading.*

# Option 3

Correction factor: $C = \left\langle \frac{\gamma}{\frac{\gamma + j}{2}} \right\rangle_j$

$$\langle p_{\mathrm{corr}} \rangle_j = \left\langle j \frac{C}{2-C} \right\rangle_j = \langle j \rangle_j \frac{\left\langle \frac{\gamma}{\gamma+j} \right\rangle_j}{1 - \left\langle \frac{\gamma}{\gamma+j} \right\rangle_j} = \langle \gamma \rangle_j \frac{k^j_{\gamma,\gamma+j} \langle j \rangle_j}{\langle \gamma \rangle_j (1 - k^j_{\gamma,\gamma+j}) + \langle j \rangle_j}$$

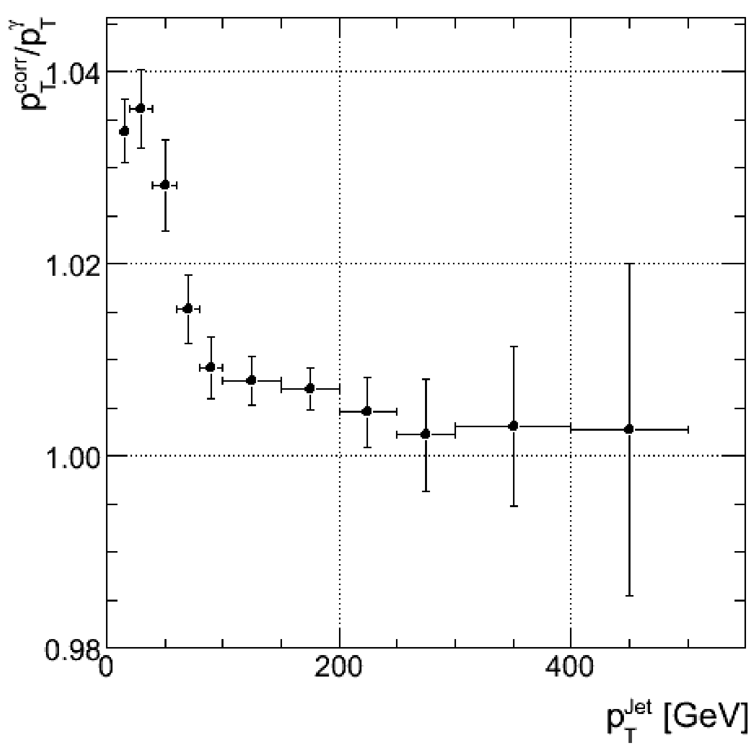


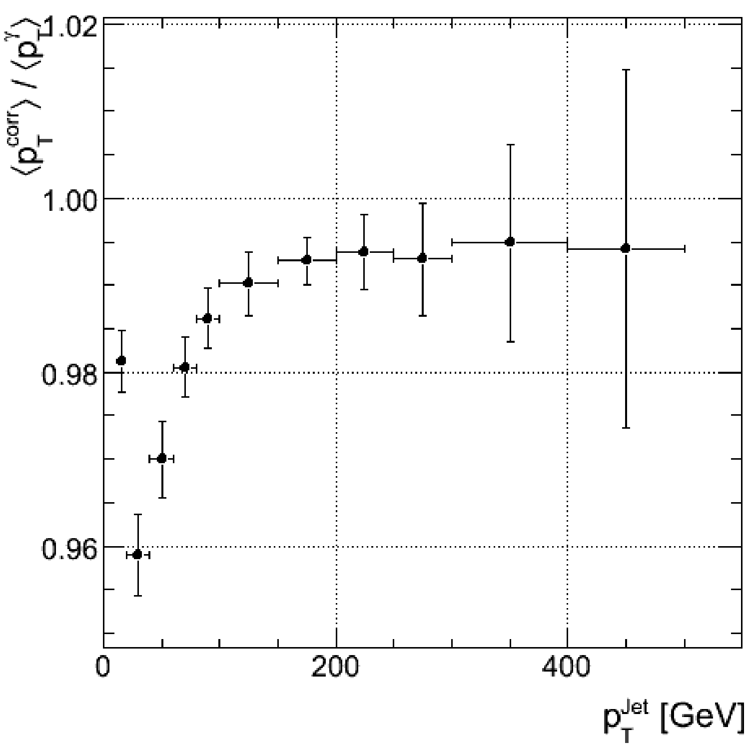
# Option 4

Correction factor: $$C = \left\langle \frac{j}{\frac{\gamma+j}{2}} \right\rangle_j$$

$$\langle p_{corr} \rangle_j = \left\langle j \frac{2-C}{C} \right\rangle_j = \langle j \rangle_j \frac{1 - \left\langle \frac{j}{\gamma+j} \right\rangle_j}{\left\langle \frac{j}{\gamma+j} \right\rangle_j} = \cdots = \langle \gamma \rangle_j + \langle j \rangle_j \left( \frac{1}{k^j_{j,\gamma+j}} - 1 \right)$$




Opt.3 & opt.4 are equivalent
(unlike opt.1 & opt.2).



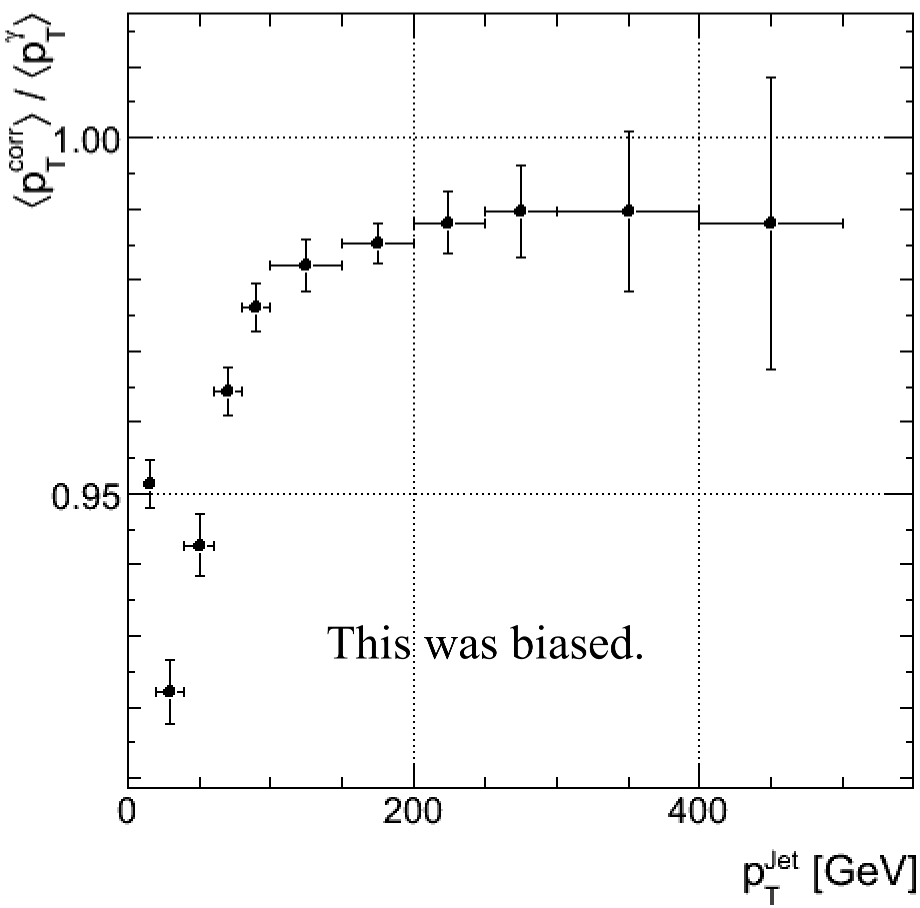
# Redefining C to remove bias

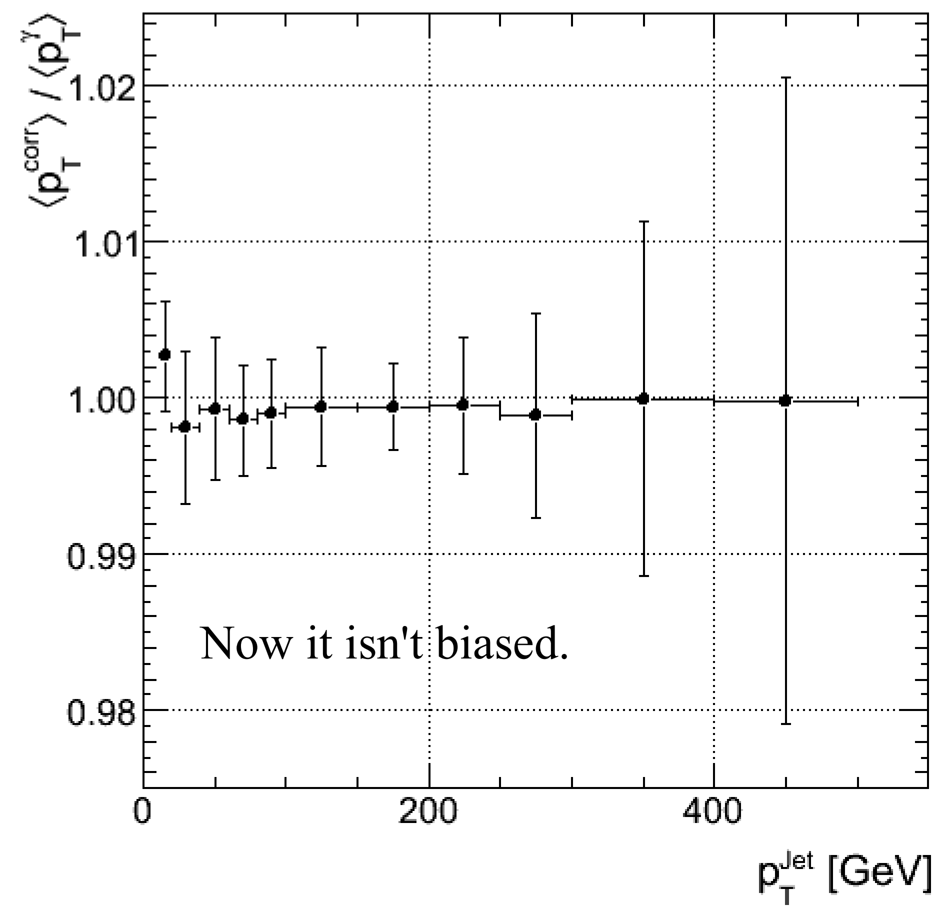
| | What people typically use | | What people should use to not bias $p_{corr}$. |
|---|---|---|---|
| • Option 1: | $C=\left\langle \frac{\gamma}{j} \right\rangle_j$ | → | $C=\frac{\langle\gamma\rangle_j}{\langle j\rangle_j}$ |
| • Option 2: | $C=\left\langle \frac{j}{\gamma} \right\rangle_j$ | → | $C=\frac{\langle j\rangle_j}{\langle\gamma\rangle_j}$ |
| • Option 3: | $C=\left\langle \frac{\gamma}{(\gamma+j)/2} \right\rangle_j$ | → | $C=\frac{2\langle\gamma\rangle_j}{\langle\gamma\rangle_j+\langle j\rangle_j}$ |
| • Option 4: | $C=\left\langle \frac{j}{(\gamma+j)/2} \right\rangle_j$ | → | $C=\frac{2\langle j\rangle_j}{\langle\gamma\rangle_j+\langle j\rangle_j}$ |

Only in this case it makes no difference.

# Option 2
# before and after unbiasing

$$C = \left\langle \frac{j}{\gamma} \right\rangle_j \quad \Rightarrow \quad C = \frac{\langle j \rangle_j}{\langle \gamma \rangle_j}$$

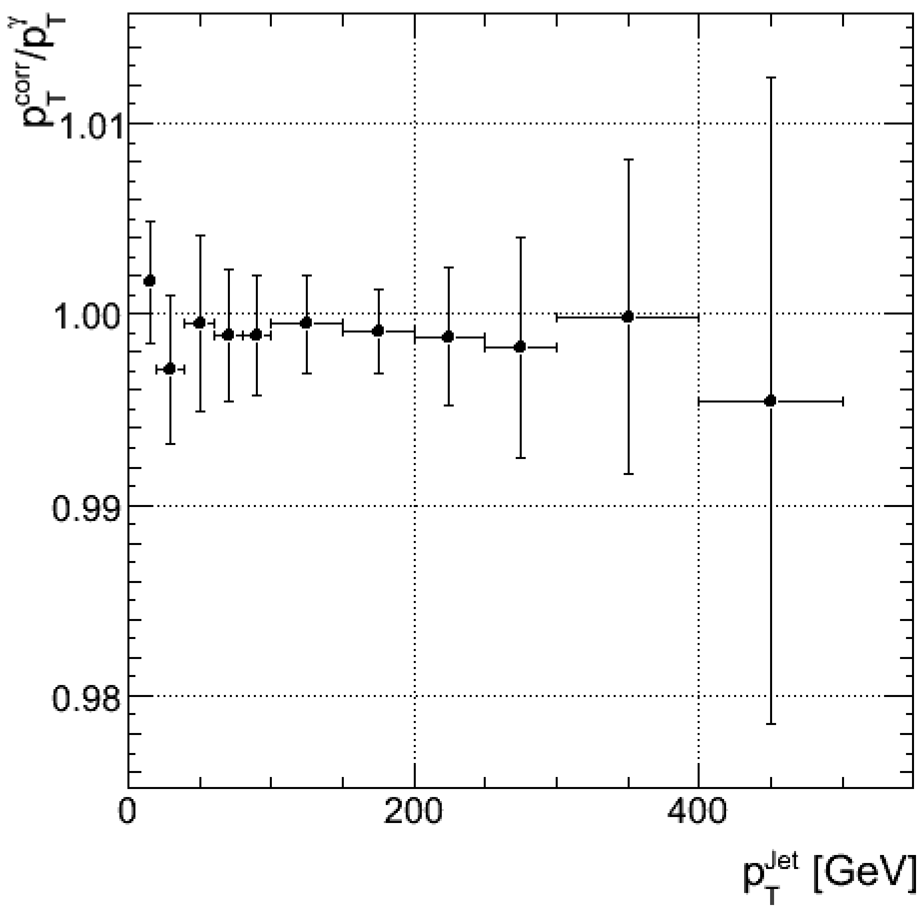


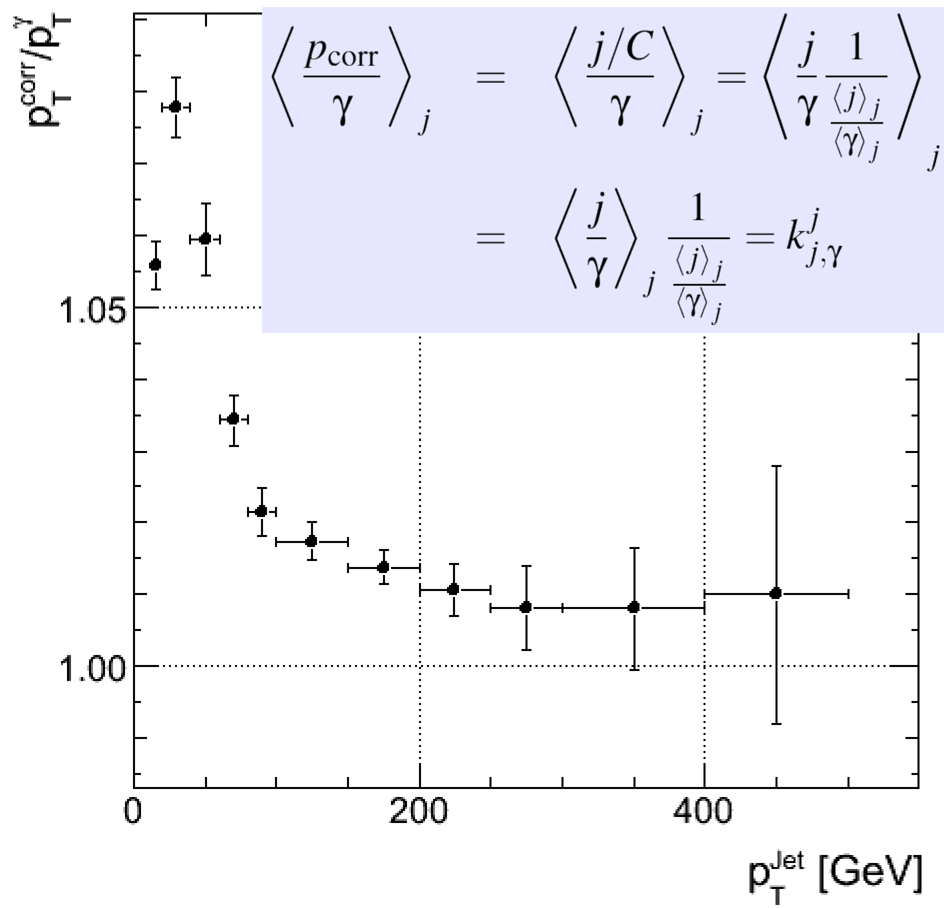



# Option 2 (continued) before and after unbiasing

In case you wonder what happened to that misleading $\langle p_{corr}/\gamma \rangle_j$ quantity.

$$C = \left\langle \frac{j}{\gamma} \right\rangle_j \quad \Rightarrow \quad C = \frac{\langle j \rangle_j}{\langle \gamma \rangle_j}$$



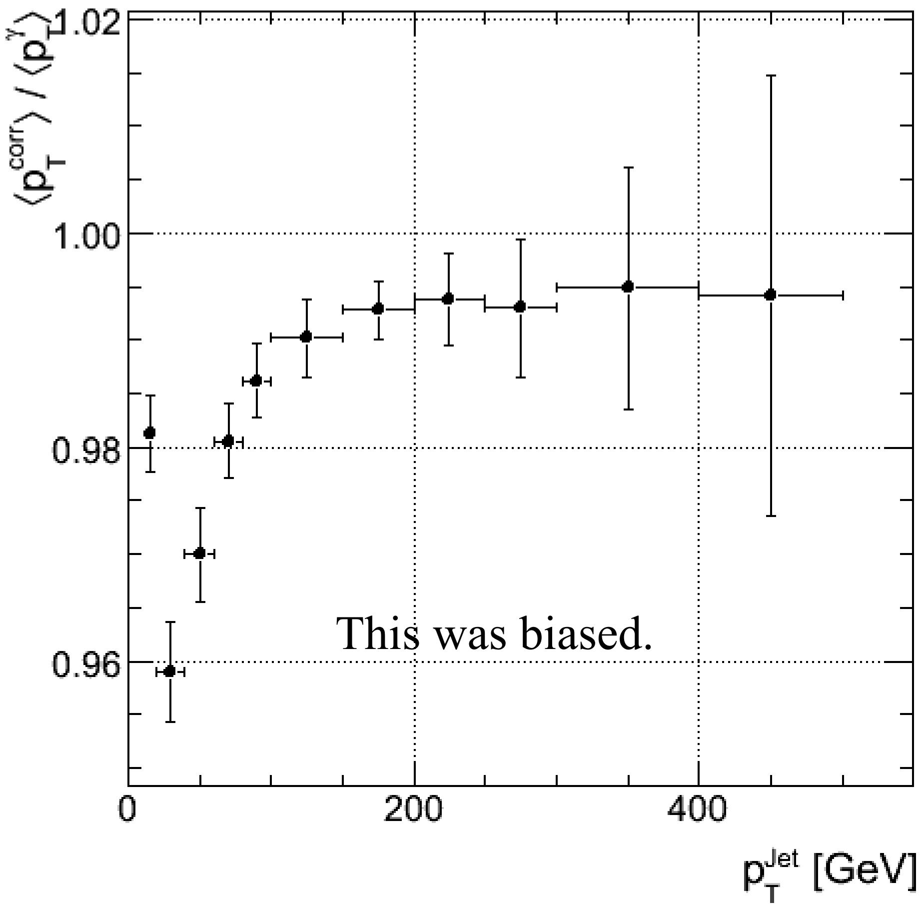



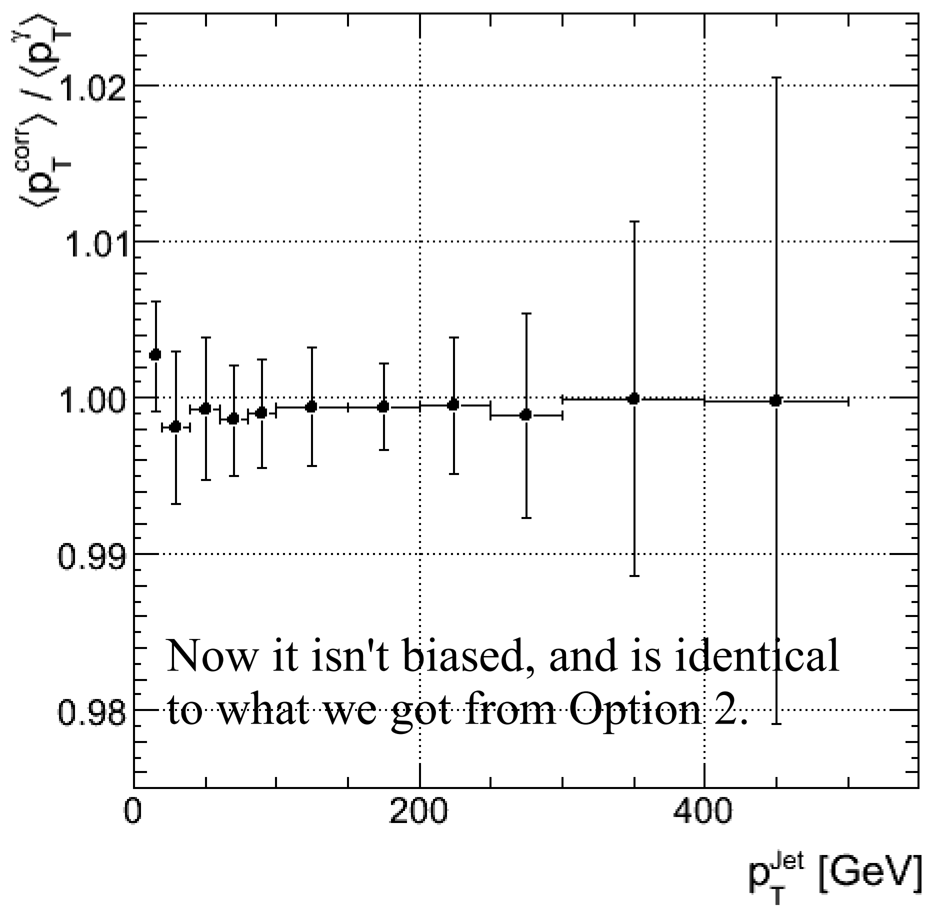

$$\left\langle \frac{p_{corr}}{\gamma} \right\rangle_j = \left\langle \frac{j/C}{\gamma} \right\rangle_j = \left\langle \frac{j}{\gamma} \frac{1}{\frac{\langle j \rangle_j}{\langle \gamma \rangle_j}} \right\rangle_j$$

$$= \left\langle \frac{j}{\gamma} \right\rangle_j \frac{1}{\frac{\langle j \rangle_j}{\langle \gamma \rangle_j}} = k^j_{j;\gamma}$$

# Option 3
# before and after unbiasing

$$C = \left\langle \frac{\gamma}{(\gamma + j)/2} \right\rangle_j \quad \Rightarrow \quad C = \frac{2\langle\gamma\rangle_j}{\langle\gamma\rangle_j + \langle j\rangle_j}$$



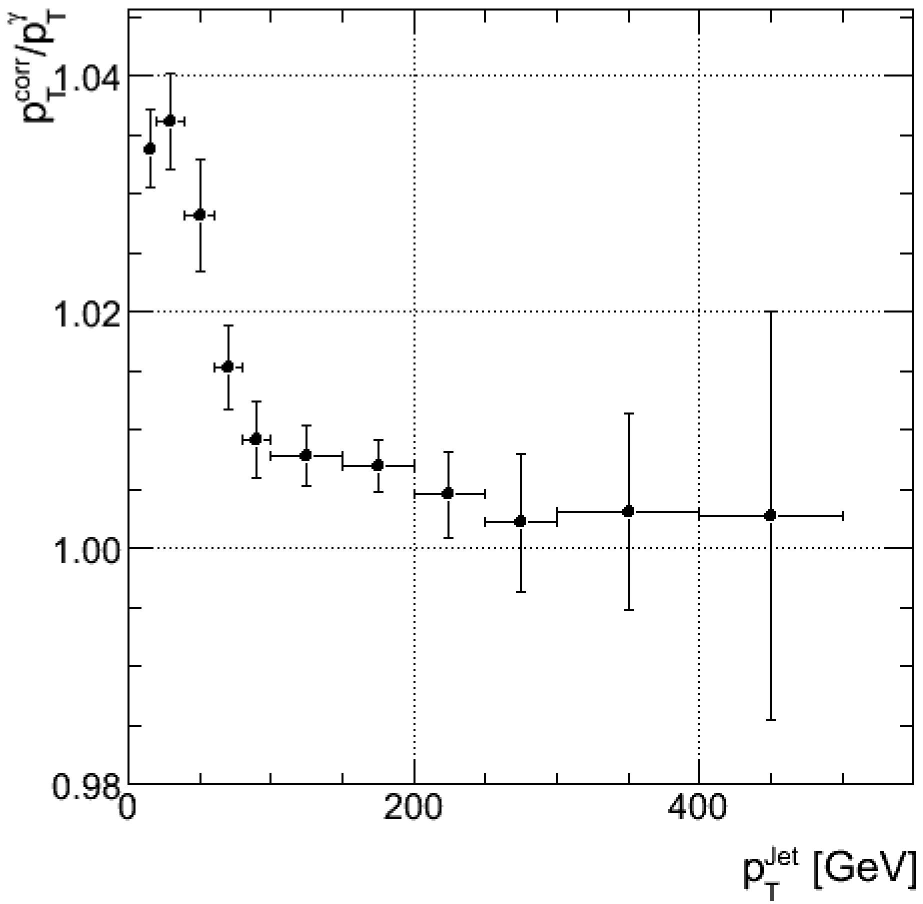



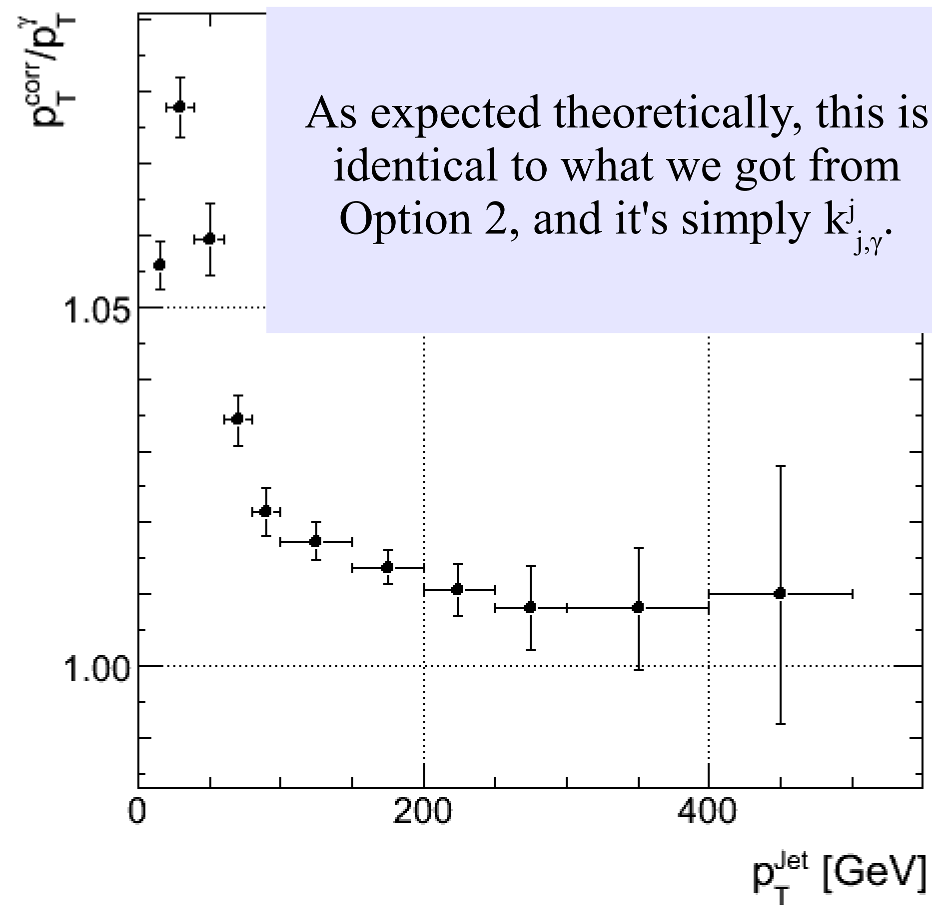
# Option 3 (continued) before and after unbiasing

In case you wonder what happened to that misleading $<p_{corr}/\gamma>_j$ quantity.

$$C = \left\langle \frac{\gamma}{(\gamma + j)/2} \right\rangle_j$$

$$C = \frac{2\langle \gamma \rangle_j}{\langle \gamma \rangle_j + \langle j \rangle_j}$$




As expected theoretically, this is identical to what we got from Option 2, and it's simply $k^j_{j,\gamma}$.

# Option 4
# before and after unbiasing

OK, you can guess...

Option 4 is identical to Option 3.

(These two were identical even before the unbiasing.)

# Summary of bias discussion so far

- We derive a correction factor C for each bin of uncorrected jet $p_T$. We can define (at least) 4 different kinds of C. Each kind needs to be applied appropriately to take us from uncorrected $p_T$ to $p_T^{corr} = p_T^{\gamma}$.

- We apply the correction in jets according to their uncorrected $p_T$. That's *the same* bin in which C was determined.

- Option 1 suffers small (but not 0) bias. The other 3 options suffer from significant bias, *due to correlation* between uncorrected jet $p_T$ and $\gamma$ $p_T$.

- We can calculate the bias analytically, and correct for it. It turns out we don't even need to measure the correlation to do that. We simply use $C=\langle j\rangle_j/\langle\gamma\rangle_j$ instead of $\langle j/\gamma\rangle_j$, for Opt. 2, and analogous transformations for the other options.

- We showed that we remove the bias, namely we get $\langle p_T^{corr}\rangle_j = \langle p_T^{\gamma}\rangle_j$ in each uncorrected jet $p_T$ bin.

- For a given option (i.e. correction definition) we may or may not see $\langle p_T^{corr} / p_T^{\gamma} \rangle_j = 1$. That is *irrelevant*. It doesn't tell us whether $\langle p_T^{corr}\rangle_j = \langle p_T^{\gamma}\rangle_j$, which is the definition of an unbiased estimator.

# Now things will get complicated

(in case they were not already)

- I showed that I can construct an estimator ($p_T^{corr}$) which is unbiased. Notice I have always been using bins defined along the variable of uncorrected jet $p_T$ (denoted by the $< >_j$ subscripts all along). I calculated C in j-bins, I applied it jet-by-jet according to jet's $p_T$, and finally I compared $<p_T^{corr}>_j$ to $<\gamma>_j$.

- But someone else wants to see what happens if we compare $<p_T^{corr}>_\gamma$ to $<\gamma>_\gamma$ for various bins of $\gamma$ (which is a short-hand for "reconstructed $\gamma$ $p_T$").

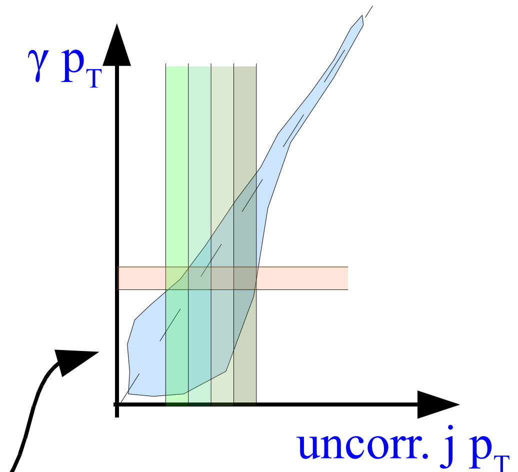
- Why does he want to see that? Because "$\gamma$ is better measured". He further claims that if $<p_T^{corr}>_\gamma \neq <\gamma>_\gamma$ then we have trouble, because "there is bias".

- The frustrating thing is that, if $<p_T^{corr}>_j = <\gamma>_j$ then $<p_T^{corr}>_\gamma \neq <\gamma>_\gamma$. In general these two conditions are *mutually exclusive*: It is impossible to be unbiased in both j-bins and $\gamma$-bins simultaneously.

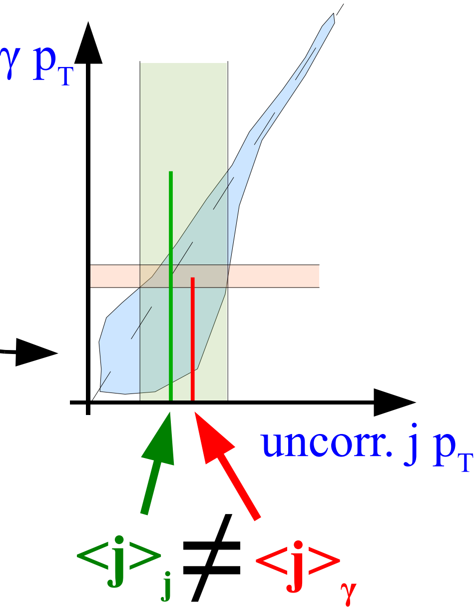
# Why is $\langle p_T^{corr} \rangle_\gamma \neq \langle \gamma \rangle_\gamma$ ?

- Take for example Option 1, that is easy and gives $\langle p_T^{corr} \rangle_j \,/\, \langle \gamma \rangle_j = k^j_{\gamma,j} = 1$.

$$\langle p_{\text{corr}} \rangle_\gamma = \langle j\, C_j \rangle_\gamma = \left\langle j \left\langle \frac{\gamma}{j} \right\rangle_j \right\rangle_\gamma$$




This already becomes troublesome: We can't pull $\langle \gamma/j \rangle_j$ out of the $\langle .. \rangle_\gamma$, because the γ-bin contains events from various j bins, hence various different correction factors $\langle \gamma/j \rangle_j$.
But OK, let's imagine that we make a j-bin so huge that it contains all the events of the γ-bin. Then, we would have just one value for $\langle \gamma/j \rangle_j$ and we could pull it outside of the $\langle .. \rangle_\gamma$:




$$\langle p_{\text{corr}} \rangle_\gamma = \left\langle \frac{\gamma}{j} \right\rangle_j \langle j \rangle_\gamma = k^j_{\gamma,j} \frac{\langle \gamma \rangle_j}{\langle j \rangle_j} \langle j \rangle_\gamma \simeq \frac{\langle \gamma \rangle_j}{\langle j \rangle_j} \langle j \rangle_\gamma$$

# What would it take to make $\langle p_T^{corr} \rangle_\gamma = \langle \gamma \rangle_\gamma$ ?

- We would have to, not only apply, but also define the correction in γ-bins. Then we would have:

$$\langle p_{\rm corr} \rangle_\gamma = \left\langle j \left\langle \frac{\gamma}{j} \right\rangle_\gamma \right\rangle_\gamma = \left\langle \frac{\gamma}{j} \right\rangle_\gamma \langle j \rangle_\gamma = k^\gamma_{\gamma,j} \frac{\langle \gamma \rangle_\gamma}{\langle j \rangle_\gamma} \langle j \rangle_\gamma = k^\gamma_{\gamma,j} \langle \gamma \rangle_\gamma$$

We see there is this $k^\gamma_{\gamma,j}$ in the end, but we already know how to deal with it. We simply redefine the correction factor to be $\langle\gamma\rangle_\gamma / \langle j \rangle_\gamma$, and that gives $\langle p_{corr} \rangle_\gamma = \langle \gamma \rangle_\gamma$.

The problem now is that this correction factor *is not applicable*.
It is defined as a function of γ $p_T$, which isn't available when we want to apply this correction to events other than γ+jet.

# My Question Is

- What do we imply when we say "unbiased estimator"? Do we mean in bins of $\gamma$ (which is well-measured but not present in QCD events), or in bins of j (which is the observable parameter of the estimator)?
- In other words, why is it "a problem" to have $\langle p_T^{corr} \rangle_\gamma \neq \langle \gamma \rangle_\gamma$ if that is the natural, well-understood result of achieving $\langle p_T^{corr} \rangle_j = \langle \gamma \rangle_j$?
- If we agree that it's required to have $\langle p_T^{corr} \rangle_\gamma = \langle \gamma \rangle_\gamma$, then we will have to define the correction factor as a function of $\gamma$. How would we apply it then?
- Let's assume there is some robust way to apply the correction which is defined as a function of $\gamma$. If that correction guarantees that $\langle p_T^{corr} \rangle_\gamma = \langle \gamma \rangle_\gamma$, then at the same time it will be making $\langle p_T^{corr} \rangle_j \neq \langle \gamma \rangle_j$. Wouldn't that be a problem? It would mean that if I apply the correction to a group of QCD jets, I would get a wrong energy, which on average would differ from the pT of the (imaginary) balancing photon. In my mind, that goes against the essence of the correction.