Top Quark Mass Measurements Using Neural Networks

Suman B. Beri[a], Pushpalatha C. Bhat[b], Rajwant Kaur[a], Harrison B. Prosper[c]

[a]Department of Physics, Panjab University, Chandigarh, India
[b]Fermi National Accelerator Laboratory, Batavia, Illinois, USA *
[c]Department of Physics, Florida State University, Tallahassee, Florida, USA

Abstract. A major goal of high energy physicists over the next few years is to reduce the uncertainty in our knowledge of the top quark mass. Neural network based methods may play a useful role in this regard, as is borne out by the preliminary results reported here.

INTRODUCTION

Over the next few years, high energy physicists hope to effect a substantial reduction in the uncertainty in our knowledge of the top quark mass (1). The success of this endeavor is contingent on the Tevatron delivering substantially more data in Run II than was collected during Run I, and, on our finding ways to reduce significantly systematic errors. As evidenced by the presentations at this workshop, neural networks (2) are extraordinarily powerful data analysis tools. We are, therefore, motivated to explore their use in the measurement of mass, with the goal of discovering methods with intrinsically smaller systematic errors. We report some preliminary results.

To provide some context we define a few pertinent concepts. A *pattern* is characterized by a d-dimensional vector $\mathbf{x}$ of real-valued numbers called *features*. Usually, one has a sample of N such *feature vectors* $\mathbf{x}_1,\ldots,\mathbf{x}_N$, each unambiguously classified into one of K classes. The feature vectors of class k inhabit a d-dimensional *feature space*, and are sampled from a probability density function $f(\mathbf{x}|k)$. An important goal of a high energy physics analysis is to assign an event of unknown identity (characterized by a vector $\mathbf{x}$) to its proper class. This can be done using the *Bayes decision rule*: Compute the (Bayesian posterior) probability $P(k|\mathbf{x})$ of $\mathbf{x}$ belonging to class k and assign $\mathbf{x}$ to the class with the highest probability. This rule minimizes the probability of making mistakes.

About a decade ago, Ruck (3), Wan (4), Blum and Li (5) showed that the outputs of a feed-forward neural network directly approximate Bayesian posterior probabilities. It is this insight that forms the basis of the methods we are studying. Although the Bayesian connection was proven in the context of neural networks the result, in fact, is much more general and pertains to *any* function able to model probability densities. Typically, such a function is obtained by minimizing an error function that has a number of adjustable parameters. It is the nature of the error function that is minimized that yields the Bayesian interpretation rather than the specific functional form that is used to model the densities. Neural networks are of interest because they are universal approximators (6) and can, therefore, in principle, model any smooth probability density.

Let $F(m,\mathbf{x},\omega)$ be a function that is able to model, to arbitrarily high accuracy, a given probability density. In the present context F is a feed-forward neural network. The quantity m is an integer in the range 1 to M, corresponding to a network with M outputs. The symbol ω denotes the set of parameters that are to be adjusted so as to minimize the error function,

$$E_N = \frac{1}{N}\sum_{n=1}^{N}\sum_{m=1}^{M}\left[D(m,\mathbf{x}_n) - F(m,\mathbf{x}_n,\omega)\right]^2 w(\mathbf{x}_n), \quad (1)$$

where D is a known function and $w(\mathbf{x})$ is an (optional) pattern dependent weight function. We choose the function D to be

$$D(m,\mathbf{x}_n) = d_{m,k} \text{ if pattern } \mathbf{x}_n \text{ belongs to class } k, \quad (2)$$

where $d_{m,k}$ is a matrix of known constants. In the limit $N \to \infty$, one obtains

$$F(m,\mathbf{x},\omega) \quad = \quad \sum_{k=1}^{K} d_{m,k} P(k|\mathbf{x}), \quad (3)$$

where the probabilities $P(k|\mathbf{x})$ are given by Bayes' theorem

$$P(k|\mathbf{x}) = \frac{w_k(\mathbf{x}) f(\mathbf{x}|k) P(k)}{\sum_k w_k(\mathbf{x}) f(\mathbf{x}|k) P(k)}, \quad (4)$$

* Operated by Universities Research Association under contract to the U.S. Department of Energy.

and where $P(k)$ is the prior probability of the kth class. We emphasize that this result holds in the limit of an infinite sample of patterns and arbitrarily accurate functional approximation. In practice, of course, neither condition holds. However, experience suggests that practical neural networks can often provide good approximations to Eq. (3), provided that the sample of patterns is large enough.

A PRELIMINARY STUDY

We consider the reaction $p\bar{p} \to t\bar{t} \to e\mu + X$, in which one top quark yields a b quark and an electron (e) and the other a b quark and a muon (μ). Although the branching ratio for the $e\mu$ channel is small, $\sim 2.4\%$ to be compared with $\sim 15\%$ for each single lepton channel, this channel is interesting because of its relatively smaller backgrounds, principally from the reactions $p\bar{p} \to Z \to \tau\tau \to e\mu$ and $p\bar{p} \to WW \to e\mu$. For this study, the $t\bar{t}$, Z and WW events were simulated using the PYTHIA program (10) and the detector was simulated using SHW (11).

The preferred way to measure the mass of a particle created in a high energy event is to perform a complete kinematic reconstruction of the event. However, for the $e\mu$ channel, as is true of the other di-lepton channels, there is insufficient information available to effect a complete kinematic reconstruction, although it is possible to perform a partial reconstruction (7). An alternative strategy is to use mass-dependent quantities that are not necessarily derived from kinematic fits (8, 9). In this study we use the variable $x(l,b) \equiv \sqrt{2l \cdot b + m_W^2}$, where m_W is the W boson mass, and l and b are lepton and b-jet four vectors. This variable, as pointed out in Ref. (8), has the virtue of being insensitive to jet systematic uncertainties, and has a distribution whose mean is an almost linear function of the top quark mass. Moreover, if the b-jet and the lepton l come from the same top quark the distribution of x ends at the top quark mass, as shown in Fig. 1a. But to avail oneself of this useful feature requires an efficient algorithm for correctly pairing the leptons and the b-jets.

In the present study we did not attempt to do this. Instead, we considered the four pairings of the two leptons and the two b-jets. As shown in Figs. 1b-e, the effect of the mispairing is to broaden the x distributions beyond the top quark mass. However, these variables are still mass-dependent and can therefore be used to measure the top mass. We associate each class label k with a top quark mass m_k; $P(k|\mathbf{x})$ may then be interpreted as the probability that the top quark mass is m_k given the measured vector $\mathbf{x} = [x(e, j_1), x(\mu, j_2), x(\mu, j_1), x(e, j_2)]$. If, furthermore, we take $M = 1$ and set $d_{1,k} = m_k$, Eq. (3) becomes the mean of the posterior probability of the top quark mass, which, from a Bayesian viewpoint, is a natural estimate of that quantity, and, moreover, is a quantity that can be directly estimated using a neural network, as noted above.

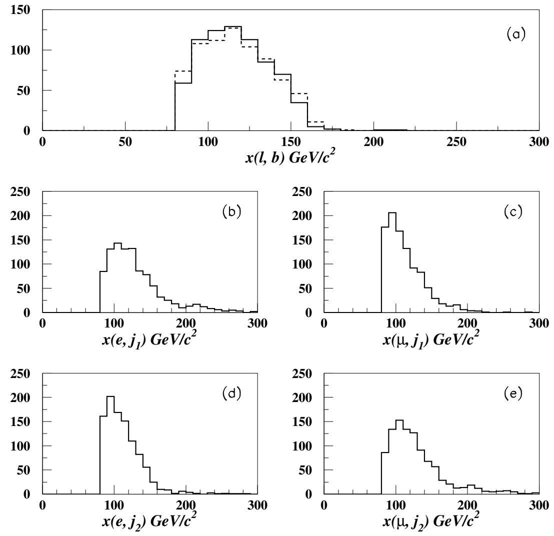

FIGURE 1. (a) The distribution of x, after event reconstruction (solid line), for correctly paired lepton and b-jet, compared with the distribution at the parton level (dashed line). We note how closely the two distributions agree. (b) to (e) The distribution of x, after reconstruction, for all pairings of leptons and jets. The distributions are for a top quark mass of 170 GeV/c^2.

To test this idea, we estimated the posterior mean using a (4,n,1) neural network (with n, the number of hidden nodes, typically equal to 5). During training we set the target value equal to the top quark mass m_k associated with the feature vector $\mathbf{x}$ (where $m_1 = 100, m_2 = 110, \ldots, m_{15} = 240$ GeV/c^2). Three hundred feature vectors were used for each top quark mass and the training was done using the program MLPfit (12) and its default minimization algorithm. To avoid over-training the network was validated on an independent sample.

Figure 2 shows the mean of the network output distribution (that is, the mean of the distribution of posterior mean) as a function of the top quark mass over the range 100 to 240 GeV/c^2. The points define a curve that is linear within the range 140 to 190 GeV/c^2 and non-linear outside. Its sigmoidal shape is easy to understand. At the lower end, the truncation at 100 GeV/c^2 and the finite width of the posterior distribution forces the posterior mean to be on average greater than 100 GeV/c^2. Likewise, at the upper end, the posterior mean is forced to be on average less than 240 GeV/c^2.

For a single measurement the method is clearly biased.

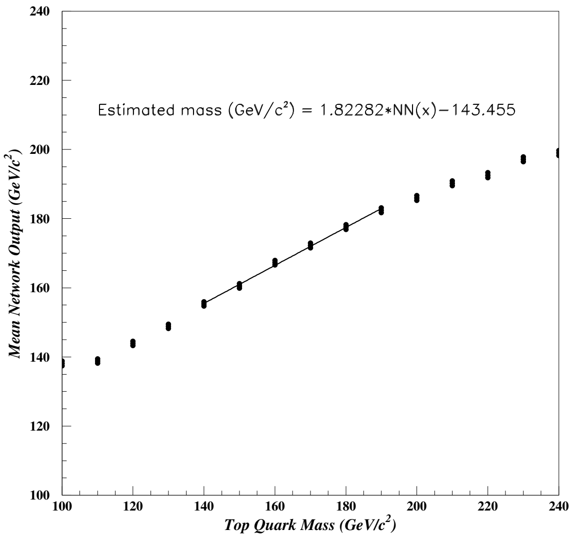

FIGURE 2. The mean of the distribution of network output as a function of the top quark mass. The solid line is the best linear fit to the specified points while the expression shown is the linear correction function derived from the fit.

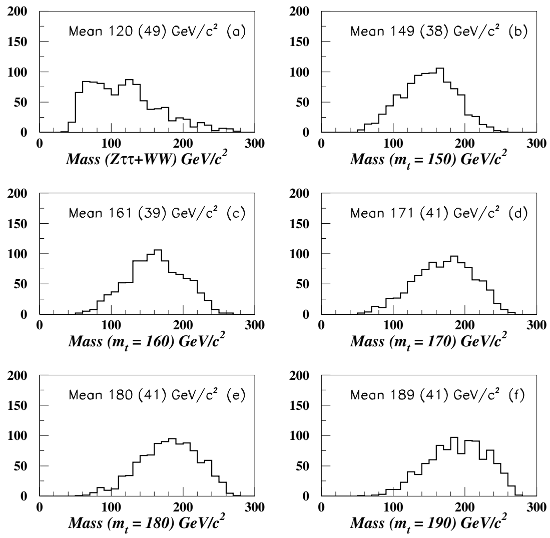

FIGURE 3. Distributions of estimated top quark mass. (a) $Z \to \tau\tau$ and WW backgrounds. (b) to (f) Signal distributions for different top quark masses. The numbers in parentheses are the estimated uncertainties for a *single* unbiased measurement, that is, for a mass measurement based on a single $e\mu$ event.

However, we argue that the relevant issue is not the bias for a sample of size one but rather the magnitude of the root mean square difference between the true mass and the measured mass for a sample whose size is likely to be much greater than one, and perhaps as large as 100. However, in this preliminary study we assume that a lack of bias for a single measurement is desirable. The simplest way to achieve this is to apply a linear correction, shown in Fig. 2, to the network output (that is, the posterior mean). When such a correction is applied one obtains the distributions shown in Fig. 3, which demonstrates the promise of the approach we have outlined.

This work was supported in part by the U.S. Department of Energy and the Department of Science and Technology of India.

REFERENCES

1. CDF and DØ Collaborations, Demortier, L. *et al.*, “The Top Averaging Group,” *Fermilab-TM-2084* (1999); Bhat, P.C., Prosper, H.B., and Snyder, S.S., *Int. J. Mod. Phys.* A **13**, 5113 (1998).
2. Bishop, C.M., *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford, 1998; DØ collaboration, Bhat, P.C, in *Proceedings of the 10th Topical Workshop on Proton-Antiproton Collider Physics*, AIP, Woodbury, New York, 1995, p. 308.
3. Ruck, D.W. *et al.*, *IEEE Trans. Neural Networks* **1 (4)**, 296 (1990).
4. Wan, E.A., *IEEE Trans. Neural Networks* **1 (4)**, 303 (1990).
5. Blum, E.K., and Li, L.K., *Neural Networks* **4**, 511 (1991).
6. Hornik, K., Stinchcombe, M., and White, H., *Neural Networks* **2**, 359 (1989); Hornik, K., *Neural Networks* **6**, 1069 (1993);
7. DØ Collaboration, Abachi, S. *et al.*, *Phys. Rev. Lett.* **80**, 2063 (1998); Kondo, K., *J. Phys. Soc. Jpn.* **57**, 4126 (1988); *ibid.*, *J. Phys. Soc. Jpn.* **60**, 836 (1991); *ibid.*, *J. Phys. Soc. Jpn.* **62**, 1177 (1993); Dalitz, R.H., and Goldstein, G.R., *Phys. Rev.* D **45**, 1531 (1992); *ibid.*, *Phys. Lett.* B **287**, 225 (1992);
8. Prosper, H.B., *Phys. Lett.* B **335**, 515 (1994).
9. CDF Collaboration, Abe, F. *et al.*, *Phys. Rev. Lett.* **80**, 2779 (1998).
10. PYTHIA v6.143, Sjöstrand, T., *Comp. Phys. Comm.* **82**, 74 (1994). Available at http://www.thep.lu.se/tf2/staff/torbjorn/Pythia.html
11. SHW v2.3, Conway, J., available at http://www.physics.rutgers.edu/~jconway/soft/shw/shw.html
12. MLPfit v1.40, Schwindling, J., and Mansoulié, B., available at http://schwind.home.cern.ch/schwind/MLPfit.html