Hunting the Higgs Boson using the Cholesky Decomposition of an Indefinite Matrix

John R. Smith and Milan Nikolic
Physics Department
University of California Davis
Davis, California 95616
and
Stephen P. Smith, Adjunct Faculty
Division of Mathematics and Science
Holy Names University
Oakland, California, 94619

January 9, 2012

Abstract

Linear models have found widespread use in statistical investigations. For every linear model there exists a matrix representation for which the ReML (Restricted Maximum Likelihood) can be constructed from the elements of the corresponding matrix. This method works in the standard manner when the covariance structure is non-singular. It can also be used in the case where the covariance structure is singular, because the method identifies particular non-stochastic linear combinations of the observations which must be constrained to zero. In order to use this method, the Cholesky decomposition has to be generalized to symmetric and *indefinite* matrices using complex arithmetic methods. This method is applied to the problem of determining the spatial size (vertex) for the Higgs Boson decay in the Higgs $\rightarrow$ 4 lepton channel. A comparison based on the χ^2 variable from the vertex fit for Higgs signal and $t\bar{t}$ background is presented and shows that the background can be greatly suppressed using the χ^2 variable. One of the major advantages of this method over the currently adopted technique of b-tagging (Tomalin, 2008) is that it is not affected by multiple interactions (pile up).

Keywords: Higgs Boson, Vertexing, b-tagging, ReML, Cholesky Algorithm, pile up, Singular Error Matrices

Submitted to the Journal of Computational and Graphical Statistics

1 Introduction

(Smith, 1995) describes how to efficiently compute both forward and backward derivatives of the Cholesky decomposition by using methods taken from automatic differentiation (Griewank, 2008). This technique permits variance-covariance estimation by restricted maximum likelihood (ReML). While the Cholesky decomposition and its derivatives are finding applications with ReML and in linear models typical to studies of animal breeding (Meyer and Smith, 1995), these are coming with additional innovations (Meyer, 2001; Meyer and Kirkpatrick, 2005). The computational methods have been introduced in general statistical packages, including SAS (SAS Institute, 2009; Meyer, 2007).

Outside of ReML, the Cholesky decomposition and its derivatives are finding the following applications: in spatial modeling or kriging (Toal, *et al.*, 2009); in lattice models with an example showing non-parametric curve fitting and cross-validation (Smith, 1997); in optimization involving the inverse diffusion problem (Christianson, 1997); to differentiate a Laplace approximation of a likelihood function, thereby permitting estimations of parameters in a population model (Frimannslund, 2006); to permit Hamiltonian Markov Chain Monte Carlo in the context of non-linear regression by function factorization (Schmidt, 2009); in calculating price sensitivities associated with exposure risk of financial portfolios (Capriotti and Giles, 2010); and for optimizing several hyper-parameters within a gradient-based machine learning algorithm (Bengio, 2000). (DeHoog, *et al.*, 2011) have introduced a general notation for treating that task involved with calculating derivatives of functions that depend on triangular matrix factors (including Cholesky's factor), and they describe several application areas too.

The above applications consider only the Cholesky decomposition of a positive-definite (or semi-definite) matrix. This convention is not followed in (Smith, 2001a and 2001b) where consideration is given to a linear state-space model, nor is it followed in the present paper. Smith used the Cholesky decomposition (and it derivatives) to estimate second moments, but in this different application the Cholesky decomposition was applied to a symmetric and indefinite matrix. In general, the Cholesky decomposition may not be possible for an indefinite matrix, in which case an attempted decomposition may lead to an interruption. Such interruption is possible when a linear model includes effects that come with a singular variance-covariance matrix structure as is possible with state-space models. It has been the case that an interrupted Cholesky decomposition

permits the correct ReML likelihood calculation following (Smith, 2001a and 2001b). However, Smith's approach implies that the linear model is consistent with the singular variance structure, and in general this assumption cannot be made. The purpose of the present paper is to get beyond this limitation by introducing linear constraints that guarantee consistency, and hence this enlarges the set of cases where Cholesky-based ReML can be applied. There are also some minor typographical errors in the pseudocode listed in (Smith, 2001a), and the corrected pseudocode is presented in the Appendix (Section 11) of this paper. Furthermore, it is the additional purpose of this paper to get beyond the common ReML applications and demonstrate a new Cholesky-based discrimination method that is suitable for enriching data samples that are used in the search for the Higgs Boson ("Higgs Hunting") at the Large Hadron Collider (LHC)

The mixed linear model and ReML are reviewed in Section 2. In Section 3, Likelihood evaluation is cast in terms of residual error, and the Cholesky decomposition of a symmetric and indefinite matrix. Section 4 provides mathematical justifications for the constraints that may be needed when the Cholesky decomposition is interrupted. To treat the possibility of constraints, Section 5 describes optimization by Lagrange multipliers and Section 6 describes optimization by penalized maximum likelihood. Both methods are readily adapted to the Cholesky decomposition and its derivatives. Section 7 treats estimation of fixed effects as an auxiliary calculation to the Cholesky decomposition by solving an indefinite system of equations. Section 8 presents an example coming from experimental physics, where the ReML method is used to devise a discrimination function to reduce backgrounds and preserve signal events in order to enrich data samples used for Higgs Hunting. The conclusion follows in Section 9.

2 Mixed Linear Models, Log-Likelihood and ReML

Notation: In the following Sections, we denote the transpose of a matrix $\mathbf{R}$ by $\mathbf{R}'$. We also denote column vector in component form by square brackets, $\mathbf{v} = [a, b, c, ..]$, and the corresponding row vectors with parentheses, $\mathbf{v}' = (a, b, c, ...)$.

Linear models, including mixed linear models, have found widespread use in statistical investigations. The linear model, though additive, is frequently flexible enough for real situations as an approximation around the mean. Also, linear models and the associated normality assumptions are well understood. Methods as old as the analysis of variance (ANOVA) are completely consis-

tent with mixed model methods. Furthermore, while the theory is developing in new areas, such the Gibbs Sampler or with other Bayesian methods, mixed model methods benefit as new tools come along. The mixed linear model is represented by:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \epsilon,$$

where $\mathbf{y}$ is a vector of observations, β is a vector of fixed effects, $\mathbf{u}$ is vector of random effects and ϵ is the observational error. The matrices $\mathbf{X}$ and $\mathbf{Z}$ are incidence matrices that relate the various effects to observations. The first moments for the random effects (their expectations) are $\mathrm{E}[\mathbf{u}] = 0$ and $\mathrm{E}[\epsilon] = 0$, and the variance-covariance structure is given by $\mathrm{var}[\mathbf{u}] = \mathbf{G}$, $\mathrm{var}[\epsilon] = \mathbf{R}$ and $\mathrm{cov}[\mathbf{u}, \epsilon] = 0$. Additional assumptions are needed to implement maximum likelihood or computer simulation, and generally $\mathbf{y}$, $\mathbf{u}$, and ϵ are taken as multivariate normal. As indicated in (Goldberger, 1962), the Best Linear Unbiased Prediction (BLUP) of $\mathbf{u}$ is found by evaluating

$$\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}[\mathbf{y} - \mathbf{X}\hat{\beta}],$$

where

$$\mathbf{V} = \mathrm{var}(\mathbf{y}) = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R} \tag{1}$$

and where $\hat{\beta}$ is the Best Linear Unbiased Estimate (BLUE) of the fixed effects obtained by the Generalized Least Squares (GSE) problem

$$\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)[\hat{\beta}] = [\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}]. \tag{2}$$

These equations can be reformulated so that the solutions can be obtained directly from the mixed model equations (Henderson, *et al.*, 1959)

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}. \tag{3}$$

However, Section 7 (see below) provides and alternative method based on the method of (Siegel, 1965).

It is well known that if a non-informative prior is used to describe the fixed effects in a Bayesian context, the posterior distribution (conditional on $\mathbf{y}$) of a linear combination of $\mathbf{b}$ (where $\mathbf{b} = [\beta', \mathbf{u}']$), say $\mathbf{H}\mathbf{b}$, is multivariate normal. In this case the posterior distribution has mean vector $\mathbf{H}\hat{\mathbf{b}}$ (where $\hat{\mathbf{b}} = [\hat{\beta}', \hat{\mathbf{u}}']$) and variance-covariance matrix given by $\mathbf{H}\mathbf{C}^{-1}\mathbf{H}'$, where $\mathbf{C}$ is the

2-by-2 partitioned matrix on the Left Hand Side (LHS) of Eq. (3). Therefore the mixed model equations are only good when the inverses $\mathbf{R^{-1}}$ and $\mathbf{G^{-1}}$ both exist. If either $\mathbf{R^{-1}}$ or $\mathbf{G^{-1}}$ does not exist, then $\mathbf{C}$ does not exist. Therefore, in the case where the covariance structure is singular, the mixed model equations will not apply.

The log-likelihood for the Multivariate Normal (MN) is given by

$$\ln(\mathbf{MN}) = \text{constant} - \frac{1}{2}\ln|\mathbf{V}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\beta)'\mathbf{V^{-1}}(\mathbf{y} - \mathbf{X}\beta)$$

The maximum likelihood estimates of β and the dispersion parameters ($\mathbf{R}$ and $\mathbf{G}$) are found by maximizing the log-likelihood. Estimates of the dispersion parameters can be badly biased by small-sample errors induced by the estimation of $\hat{\beta}$. This is a serious problem when the dimension of β is large relative to the information available to estimate β.

To overcome this problem (Patterson and Thompson, 1971) introduced Restricted Maximum Likelihood (ReML), where the dispersion parameters are found by maximizing

$$\ln(\mathbf{ReML}) = \text{constant} - \frac{1}{2}\ln|\mathbf{V}| - \frac{1}{2}\ln|\mathbf{X'VX}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\hat{\beta})'\mathbf{V^{-1}}(\mathbf{y} - \mathbf{X}\hat{\beta}), \tag{4}$$

where $\hat{\beta}$ is the solution obtained by GSE, Eq. (2). ReML has the advantage of estimating away the β parameters from the Likelihood. This is especially useful in cases where one wants to concentrate on minimizing the deviations from a common mean, without explicitly finding that common mean. (Wikipedia, 2011) states, "In contrast to conventional maximum likelihood estimation, ReML can produce unbiased estimates of variance and covariance parameters." (Harville, 1974) derived the likelihood in Eq. (4) to treat the "error contrasts" which are found by taking a complete set of linear combinations of the observations which are sufficient to remove the effect of β while leaving the maximal amount of information for the purpose of ReML. An early review of ReML can be found in (Harville, 1977). More reviews can be found in (Speed, 1977 and 1995). The relevance for the particular application described in Section 8 (see below), is that the χ^2 that is determined by ReML is independent of the central vertex coordinates. It is possible to back-out the coordinates of the fitted vertex, but it is not necessary to know the coordinates in order to determine the goodness of fit.

3 Applying ReML to the Z = 0 Scenario

Consider the following linear model $\mathcal{L}$ in the case where $\mathbf{Z} = \mathbf{0}$

$$\mathcal{L} : \mathbf{y} = \mathbf{X}\beta + \epsilon,$$

where $\mathbf{y}$ is an independent observation vector, β is a vector of fixed effects which are to be removed using the ReML likelihood, and ϵ is a vector of random residuals. The array $\mathbf{X}$ is the incidence matrix that assigns fixed effects to observations. The variance-covariance matrix of the random residuals is denoted by $\mathbf{R}$

$$\text{var}[\epsilon] = \mathbf{R}.$$

The linear model $\mathcal{L}$ is now fully specified and the error matrix satisfies $\mathbf{V} = \mathbf{R}$ so that the likelihood given by Eq. (4) reduces to

$$\ln(\mathbf{ReML}) \rightarrow \text{constant} - \frac{\mathbf{1}}{\mathbf{2}} \ln |\mathbf{R}| - \frac{\mathbf{1}}{\mathbf{2}} \ln |\mathbf{X'RX}| - \frac{\mathbf{1}}{\mathbf{2}} (\mathbf{y} - \mathbf{X}\hat{\beta})' \mathbf{R^{-1}} (\mathbf{y} - \mathbf{X}\hat{\beta}). \tag{5}$$

(Smith, 2001b) associates a symmetric matrix $\mathbf{K}$ with the above linear model $\mathcal{L}$ as follows:

$$\mathbf{K} = \begin{pmatrix} \mathbf{R} & \mathbf{X} & \mathbf{y} \\ \mathbf{X}' & 0 & 0 \\ \mathbf{y}' & 0 & 0 \end{pmatrix}, \tag{6}$$

where the 0's denote the appropriate-sized null square matrices required to fill out the rows and columns of $\mathbf{K}$ and $\mathbf{X}'$ is the transpose of $\mathbf{X}$. Therefore, $\mathcal{L}$ implies the existence of a matrix $\mathbf{K}$, and $\mathbf{K}$ implies there exists a model $\mathcal{L}$. Our main interest in Eq. (6) is that $\mathbf{R}$ is *not required* to be invertible. The method we use is able to identify those linear combination of $\mathbf{y}$ which are associated with the zero-eigenvalues of $\mathbf{R}$. These linear combinations are non-stochastic and can be eliminated from the stochastic part of the likelihood and treated by the method of Lagrange constraints. In other words, our method is able to find the natural constraints required for the maximum likelihood problem for ReML.

It is well known that the Cholesky decomposition runs to completion with any matrix that is symmetric and non-negative definite. (Smith, 2001b) shows that the Cholesky decomposition can also be performed on the matrix $\mathbf{K}$, and this is most curious because while $\mathbf{K}$ is symmetric, it is not non-negative definite. $\mathbf{K}$ is classified as indefinite. However, the rows and columns of

$\mathbf{K}$ must be first permuted, leaving the last row and column in place as required. This is enough to permit computation of the Cholesky decomposition, or the lower triangular matrix $\mathbf{L}$, where $\mathbf{K_{k\times k}} = \mathbf{LL}'$ (for some permutation involving the first $k-1$ rows and columns). (Smith, 2000 and 2001a) describes the ReML function, or the likelihood that removes the impacts of β, to be a function of this particular $\mathbf{L}$. This calculation is given neatly as follows:

$$\ln(\mathbf{ReML}) \rightarrow \text{constant} - \sum_{i<k} \ln|L_{i,i}| + \frac{1}{2}L_{k,k}^2,$$

where the absolute modulus function $|L_{i,i}|$ transforms a possible imaginary number into a positive number, and the summation is only over the non-zero pivots. When zero pivots are encountered (when $L_{i,i} = 0$ for some i) *we require that the i-th column of* $\mathbf{L}$ *vanishes and this is enough to justify likelihood evaluation by the above formula*. It is easy to show that this calculation agrees with the likelihood given in Eq. (5) when R is non-singular. However, in the most general case we cannot simply skip the zero pivots, and more will be said about this in the Section 4, because particular constraints must be imposed.

The likelihood function is derived from elements of the Cholesky decomposition, and so there is nothing else that is needed to perform ReML but to find the derivatives that permit the likelihood to be optimized by the iterative Newton-Raphson technique. These derivatives come automatically with the Cholesky decomposition (Smith, 2000 and 2001a), and so there is little beyond the matrix $\mathbf{K}$, and its decomposition, that must be considered to describe ReML. The corrected pseudocode for the differentiation of the Cholesky algorithm and directions for how to use it are found in the Appendix (see Section 11).

4 Justification and Additional Constraints

The Cholesky decomposition of a matrix $\mathbf{K_{k\times k}}$ proceeds by identifying a non-zero diagonal element (the first pivot), and then permuting rows and columns of $\mathbf{K}$ to reposition that diagonal to the first diagonal position. Elementary row operations are now conducted to annihilate elements below the diagonal in the first column (this is pivoting). The first row of $\mathbf{K}$ is now transformed into the first column of $\mathbf{L}$ by replacing the first diagonal by its square root and by dividing the remaining elements in the first row by that square root. The first row and column of $\mathbf{K}$ are now deleted to produce a smaller sub-matrix of order $k-1$. This sub-matrix is called the Schur complement. An

outline computation of the first Schur complement (of K_{11} in $\mathbf{K}$) is displayed below.

$$\begin{pmatrix} K_{11} & K_{12} & ... & K_{1k} \\ K_{21} & K_{22} & ... & K_{2k} \\ ... & ... & ... & .. \\ K_{1k} & K_{k2} & ... & K_{kk} \end{pmatrix} \rightarrow \begin{pmatrix} K_{22} & K_{23} & ... & K_{2k} \\ K_{32} & K_{33} & ... & K_{3k} \\ ... & ... & ... & ... \\ K_{k2} & K_{k3} & ... & K_{kk} \end{pmatrix} - \frac{1}{K_{11}} \begin{pmatrix} K_{21} \\ K_{31} \\ ... \\ K_{k1} \end{pmatrix} (K_{12}, K_{13}, ..., K_{1k})$$

When $\mathbf{K}$ is symmetric, the Schur complement is symmetric. Therefore, the Cholesky decomposition may proceed by half-storing $\mathbf{K}$. The above steps are now repeated to generate a second Schur complement, then a third and so on. The algorithm may be organized to overwrite the half-stored $\mathbf{K}$ with $\mathbf{L}$. The operations of the Cholesky decomposition are reversible, and therefore, the Cholesky decomposition conserves information.

(Smith, 2001b) generalized the Cholesky decomposition for the case when $\mathbf{K}$ is symmetric and indefinite by using a complex representation for the Cholesky diagonals if needed. The sub-matrix initially containing the zero entries is the non-positive definite partition, and with fill-in (generated from pivoting from the diagonals of $\mathbf{R}$) the non-positive definite partition becomes more negative. When pivoting switches over to the diagonals of the non-positive definite partition, then fill-in results in the partition initially set to $\mathbf{R}$ (or the non-negative definite partition) thereby making it more positive. If a few pivots are selected from the diagonals of the non-negative definite partition first, before switching over to the non-positive partition and continuing pivoting over the negative diagonals until the non-positive definite partition returns to a matrix containing zero in all its entries (expect the last diagonal), then (Smith, 2001b) referred to the particular pivot order as a *standard data reduction.*

After the first pivot step within the Cholesky decomposition, the lead row and column is removed from $\mathbf{K}$, and this reduces the dimension of the resulting Schur complement by 1 as noted above. After a standard data reduction, the Schur complement is again reduced in dimension from $\mathbf{K}$ but retains the special matrix form (ignoring the last diagonal with no loss in generality):

$$\mathbf{K_1} = \begin{pmatrix} \mathbf{R_1} & \mathbf{X_1} & \mathbf{y_1} \\ \mathbf{X_1'} & 0 & 0 \\ \mathbf{y_1'} & 0 & 0 \end{pmatrix},$$

where the subscript indicates that $\mathbf{K_1}$ was generated from $\mathbf{K}$ following the Cholesky decomposition. We will denote $\mathbf{K}$ as $\mathbf{K_0}$, to maintain consistency. In shorthand, this transformation is denoted

by $\mathbf{K_0} \rightarrow \mathbf{K_1}$.

(Smith, 2001b) notes that $\mathbf{K_1}$ also represents a model $\mathcal{L}_1$, given by

$$\mathcal{L}_\mathbf{1} : \mathbf{y_1} = \mathbf{X_1}\beta_\mathbf{1} + \epsilon_\mathbf{1},$$

where $\beta_\mathbf{1}$ is sub-vector of β and the variance-covariance matrix of $\epsilon_\mathbf{1}$ is $\mathbf{R_1}$. The model $\mathcal{L}_1$ is fully determined given the Schur complement $\mathbf{K_1}$. Moreover, those observations that have been processed and eliminated from $\mathbf{K}$ are uncorrelated with $\epsilon_\mathbf{1}$. Also, the extracted information was accounted for in the previously constructed Cholesky diagonal elements. In other words, the standard data reduction separates the data into two statistically independent parts. The first statistically independent part is used to compute that block of the log-likelihood already known for ReML and consistent with Eq. (5). The remaining part pertains to the treatment of $\mathcal{L}_1$, but because $\mathbf{K_1}$ is in the form of $\mathbf{K}$, the process can be iterated and the Cholesky decomposition continued to the next pivot.

The Cholesky decomposition signifies a tower of standard data reductions:

$$\mathbf{K_0} \rightarrow \mathbf{K_1} \rightarrow \mathbf{K_2} \rightarrow \mathbf{K_3} \rightarrow \mathbf{K_4}$$

And these Schur complements correspond to a tower of models:

$$\mathcal{L}_0 \rightarrow \mathcal{L}_1 \rightarrow \mathcal{L}_2 \rightarrow \mathcal{L}_3 \rightarrow \mathcal{L}_4$$

At each transition the dimensions of the Schur complement get smaller and smaller, and all along the way information is being processed correctly to evaluate the ReML likelihood. The only question is whether the Cholesky decomposition completes and leads to a final Schur complement that folds into the likelihood function calculation. However, the last Schur complement may be one of two special forms which cannot be reduced further. These are the First Special Form:

$$\begin{pmatrix} \mathbf{0} & \mathbf{H} & \mathbf{v} \\ \mathbf{H}' & \mathbf{0} & \mathbf{0} \\ \mathbf{v}' & \mathbf{0} & \mathbf{p} \end{pmatrix},$$

or the Second Special Form:

$$\begin{pmatrix} \mathbf{0} & \mathbf{v} \\ \mathbf{v}' & \mathbf{p} \end{pmatrix},$$

In the case of an incomplete Cholesky decomposition, what is made manifest are the constraints that must also be imposed on the maximization of the likelihood function. The remarkable conclusion is that all the information is contained within the Cholesky decomposition (as we will see), even when the Cholesky decomposition is unable to finish.

One might question the appropriateness of the standard data reduction, given that the pivot order in the Cholesky decomposition may be dynamic and may not follow the standard data reduction. However, (Smith, 2001b) proved that pivot orders come in equivalence classes. Any dynamic order corresponds to a tower of standard data reductions where the ReML likelihood is treated correctly. And moreover, Schur complements are invariant to the pivot order that generates them (including the last diagonal of the Cholesky decomposition), as well as the determinant calculated as the product of pivots. The correct likelihood is calculated even when the standard data reduction is not followed, because implicit in any pivot order is a tower of standard data reductions.

With dynamical pivoting, the Cholesky decomposition is permitted to go as far as it can while skipping zero diagonals that would otherwise be pivots. Some zero pivots may encounter fill-in during the computation, and this permits the Cholesky decomposition to continue with additional rounds of pivoting. If for some reason the Cholesky decomposition in unable to complete the operations, the matrix that remains (as unfinished) will be a sub-matrix of what had been $\mathbf{K}$ and what was becoming $\mathbf{L}$ (but never completed). The unfinished sub-matrix contains the last Schur complement either in the First or Second Special Forms (noted above), even if the pivot order did not follow a tower of standard data reductions.

If extraneous parameters remain impacting the likelihood function, they are found involved in non-stochastic linear combinations, and these combinations are revealed in the Schur complement (the First Special Form containing both $\mathbf{H}$ and $\mathbf{v}$) that could not be reduced by the Cholesky decomposition:

$$\mathbf{H}\beta_{\mathbf{I}} = \mathbf{v},$$

where $\beta_{\mathbf{I}}$ is a sub-vector of β, and $\mathbf{v}$ is a revealed linear combination of $\mathbf{y}$. If the extraneous parameters are no longer present, then the matrix $\mathbf{H}$ goes away. We are left with the Second Special Form of the Schur complement that only involves $\mathbf{v}$ equated to a column vector of zeros:

$$\mathbf{v} = \mathbf{0}.$$

When the Cholesky decomposition is unable to finish, then one of these systems of linear equations becomes a side condition (a constraint) for the likelihood maximization exercise.

We may be uninterested in the extraneous parameters. However, we are interested in consistent models that don't contradict themselves when linear combinations are made of their components. This concern is quite independent of the extraneous parameters as we will see. The set of non-stochastic equations that are revealed (by the First Special Form of the Schur complement) will involve extraneous parameters, and these will be ignored in as much as ReML removes their impacts from the likelihood. The revealed set of equations can imply whatever they want about $\beta_{\mathbf{I}}$ and ReML will ignore them. This assumes that the statistical model is already consistent. However, the non-extraneous parameters that are identified for likelihood evaluation need not conform to a consistent linear model and that's where we depart from (Smith, 2001b).

Likewise, if the attempt at Cholesky decomposition ends with a Schur Complement of the Second Special Form (leading to the equation $\mathbf{v} = \mathbf{0}$), what is revealed are constraints that must be imposed to guarantee a consistent linear model. In this case there are no extraneous parameters to confuse the issue.

5 Constraints: Lagrange Multiplier Method

In the event that the Cholesky decomposition encounters a zero and is unable to complete, in order to to maintain consistency one can consider appending to the ReML likelihood a Lagrange multiplier expression according to one for the following cases:

$$\text{Case 1}: \quad F_1(\mathbf{L}, \lambda) = \text{constant} - \sum_{i<k} \ln |L_{i,i}| + \frac{1}{2} L_{k,k}^2 + \lambda'(\mathbf{H}\beta_{\mathbf{I}} - \mathbf{v}),$$

where maximization proceeds with respect to the non-extraneous parameters (s_1, s_2, s_3, s_4), $\beta_{\mathbf{I}}$ and λ. Therefore, the side condition enforces the restriction that $\mathbf{v}$ is in the column space of $\mathbf{H}$ independent of $\beta_{\mathbf{I}}$.

$$\text{Case 2}: \quad F_2(\mathbf{L}, \lambda) = \text{constant} - \sum_{i<k} \ln |L_{i,i}| + \frac{1}{2} L_{k,k}^2 + \lambda'\mathbf{v},$$

where maximization proceeds with respect to the non-extraneous parameters (s_1, s_2, s_3, s_4) and λ. Therefore, the side conditions enforces the restriction that $\mathbf{v} = \mathbf{0}$.

In both cases, we find an objective function so constructed from the known elements of the unfinished Cholesky decomposition. First and second derivatives are available (Smith, 2001a) and constrained optimization is straightforward by the technique of Newton-Raphson iteration as applied in Section 8 below.

In paradoxical cases, the data may produce an inconsistent model. In which case, we recommend constrained maximization as noted above, and in this way useful information is included that would otherwise be ignored.

In case the Cholesky decomposition is unable to finish and the Schur complement is of the Second Special Form, then the particular linear combination of elements corresponding to $\mathbf{v}$ can be constrained to zero by adding a Lagrange multiplier term $\lambda\mathbf{v}$ as indicated above. An alternative method involves adding a quadratic penalty term to the likelihood (see Section 6) can be used to adjust the constraints to zero within estimated precision.

Case 1 and Case 2 above represent possible objective functions. However, we concentrate on Case 2 because that is the one relevant to our example. The objective function to be maximized is of the form $F_2(\mathbf{L}, \lambda)$. The three algorithms of Section 11 are used to construct the Newton-Raphson linear system. If required, the first derivatives with respect to the Lagrange multipliers (the λ vector) come directly out of the $\mathbf{v}$ elements of the corresponding Schur complement of the unfinished matrix $\mathbf{L}$. Table 1 sketches the Cholesky decomposition which is used to calculate the components of the ReML likelihood and the possible non-stochastic linear combinations that will be subject to the constraint conditions. We initialize the array $\mathbf{F}$ of Table 2 to the array $\partial F_2(\mathbf{L}, \lambda)/\partial L_{ij}$ that represents the constrained objective function including the Lagrange multipliers. We then find the mixed second derivatives directly from $\mathbf{Q}$ (see Table 3, which represents the forward derivative calculations.)

6 Constraints: Penalized Likelihood Method

There is another method of implementing constraints besides Lagrange multipliers. Instead of introducing a set of Lagrange multipliers which are treated as independent parameters to be varied in a maximum likelihood procedure, the terms which they multiply can be squared and then multiplied by a fixed constant. This is known as the "Penalized Likelihood" method. In the Penalized Likelihood method the Lagrange equations for Case 1 and Case 2 above are modified

to the form

$$\text{Case 1}: \quad F_1(\mathbf{L}) = \text{constant} - \sum_{i<k} \ln |L_{i,i}| + \frac{1}{2} L_{k,k}^2 + \sum \mathbf{c_i} (\mathbf{H}\beta_\mathbf{I} - \mathbf{v})_\mathbf{i}^\mathbf{2},$$

where the c_i are nominated constants and maximization proceeds with with respect to the non-extraneous parameters (s_1, s_2, s_3, s_4) and $\beta_\mathbf{I}$.

$$\text{Case 2}: \quad F_2(\mathbf{L}) = \text{constant} - \sum_{i<k} \ln |L_{i,i}| + \frac{1}{2} L_{k,k}^2 + \sum \mathbf{c_i} (\mathbf{v_i})^\mathbf{2},$$

where c_i are nominated constants and maximization proceeds with respect to the non-extraneous parameters (s_1, s_2, s_3, s_4). The squares of the column vectors in the last term on the RHS of the above are to be understood as vector dot products.

The main difference between the Lagrange Multiplier method and the Penalized Likelihood method is that the auxiliary constraints are imposed to the maximum machine precision by the Lagrange Multiplier method, whereas the constraints are imposed in a less extreme manner with the Penalized Likelihood method. The point of the nominated constants is to adjust the enforcement of the constraints to within reasonable limits set by the known precision of the measurements. This allows the possibility, for example, of not demanding that the constraints be satisfied with more rigor than the measurement uncertainties can justify.

7 Estimating the Fixed Effects $\hat{\beta}$

(Siegel, 1965) described a method to compute generalized least squares (GLS) estimates by way of a system of equations with a symmetric and indefinite coefficient matrix. In our notation, this system of equations is presented below:

$$\begin{pmatrix} \mathbf{R} & \mathbf{X} \\ \mathbf{X}' & \mathbf{0} \end{pmatrix} \begin{pmatrix} \lambda \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix},$$

where $\hat{\beta}$ is the GLS estimate of β, and λ are the Lagrange multipliers. This Section will follow (Smith, 2001a), and show how to calculate both $\hat{\beta}$ and λ as adjunct operations that depict backward substitution, given that the Cholesky decomposition of $\mathbf{K}$ is available.

Note that the both the coefficient matrix and right-hand side of Siegel's equations are sub-matrices of $\mathbf{K}$: the coefficient matrix is the lead sub-matrix in $\mathbf{K}$, and the right-hand side is fixed to the last column (or row) and remains there for all permitted row-column permutations of $\mathbf{K}$.

Now rewrite Siegel's equations in simple terms:

$$\mathbf{Cb} = \mathbf{r},$$

where $\mathbf{C}$ and $\mathbf{r}$ signify the coefficient matrix and right-hand side, respectively; and $\mathbf{b}$ is a column vector containing λ and $\hat{\beta}$. To solve these equations, we might permute the rows and columns of $\mathbf{C}$, and compute the Cholesky decomposition: $\mathbf{TT}' = \mathbf{C}$. To permute the rows and columns of $\mathbf{C}$, and the rows of $\mathbf{b}$ and $\mathbf{r}$ are also permuted to leave Siegel's equations intact. Now multiplying both sides of Siegel's equations by the same elementary row operations that transform $\mathbf{C}$ into $\mathbf{T}'$ gives:

$$\mathbf{T}'\mathbf{b} = \mathbf{a}.$$

When $\mathbf{C}$ is non-singular, $\mathbf{a} = \mathbf{T}^{-1}\mathbf{r}$. The coefficient matrix in this new system of equations is upper triangular. Therefore, vector $\mathbf{b}$ can be solved by backward substitution.

With $\mathbf{L}$ computed, where $\mathbf{K} = \mathbf{LL}'$, note that $\mathbf{T}$ is the leading sub-matrix in $\mathbf{L}$ and $\mathbf{a}'$ is the last row vector of $\mathbf{L}$ (excluding the last diagonal). Therefore, having computed $\mathbf{L}$ we need only enter backward substitution to evaluate $\mathbf{b}$. The GLS estimates $\hat{\beta}$ will be found scattered in $\mathbf{b}$, noting that $\mathbf{b}$ is permuted.

When the i-th pivot encountered in $\mathbf{L}$ is zero, and the i-th column vanishes, a singularity is present and $\mathbf{b}$ is not estimated uniquely. We place a restriction on $\mathbf{b}$: set its i-th entry to zero. This modification is implemented when the Cholesky decomposition is unable to finish but ends with a Schur complement of the Second Special Form. The associated column of $\mathbf{L}$ will also be constrained to vanish. When the Cholesky decomposition ends with the First Special Form then an auxiliary estimate of $\beta_{\mathbf{I}}$ is available from the main optimization and this estimate is used in backward substitution given that what had been computed for the rest of $\mathbf{L}$ is lower triangular.

8 Example: Spatial Errors in Track Extrapolations and Vertexing in Higgs $\rightarrow$ 4 lepton Searches

One of the most important decay modes of the Higgs Boson is into 4 charged leptons (Higgs $\rightarrow$ 4 leptons). In a solenoidal detector there is a large magnetic field present which is represented as a constant vector, $\vec{\mathbf{B}}$, which is taken by convention to point in the z-direction. Each of the charged leptons will approximately trace out a right-circular (or left-circular) helix with symmetry axis also parallel to the z-direction. Tracks undergo multiple scattering and energy losses as they traverse the detector which limit the accuracy of the helical path assumption. These effects are usually very small for tracks in the central solenoid region that have high transverse momentum p_T (momentum component perpendicular to the z-axis). The radius of a track's helical path depends its p_T. Each helix in 3-d has the form $\vec{r}(s) = [x(s), y(s), z(s)]$ and is a function of an independent position parameter s that marks the location along the track path. The components of the position vector $\vec{r}(s)$ are given by

$$\begin{aligned} x(s) &= (\rho - dxy)\sin\phi - \rho\sin(\phi - s/\rho) \\ y(s) &= \rho\cos(\phi - s/\rho) - (\rho - dxy)\cos\phi \\ z(s) &= Z + s\cot\theta \end{aligned} \tag{7}$$

The parameters on the RHS of Eq. (7) are obtained from the track-fitting algorithm and have the following meaning:

q = charge of particle

k = curvature of track circle, R = 1/k = radius of curvature

$Z =$ z-coordinate of the point on the track helix closest to z-axis

θ = polar angle of the tangent to track at Z

ϕ = azimuthal coordinate ($\tan\phi = -$x/y) of the track helix at Z

$\rho = qR$

$D =$ signed distance from beam axis to track helix at Z

$dxy = qD$

$dsz = Z/\sin(\theta)$

where the sign of the distance for D is given as positive if the track circle (projection of helix into the x-y plane) contains the z-axis and negative otherwise. A typical track fitting program will produce the 5-parameters $(\rho, \phi, \theta, dxy, dsz)$ along with a 5-by-5 covariance matrix which can be used to estimate the spatial error matrix in terms of the 3-by-3 correlation matrix for $x(s), y(s), z(s)$, for example, by the method of "propagation of errors". Once these track parameters are measured, then the track can be extrapolated to any position along its trajectory by using the parametric equation of a helix as a function of s. One of the interesting properties of these spatial extrapolations is the the 3-by-3 spatial error matrix is very close to rank-2 (almost singular). Fig. (1) illustrates the reason that the spatial error matrix has a near-zero eigenvector lying almost entirely in the x-y plane. If the radius of the track circle is much larger than the radius of the tracking detector, then only a small fraction of the circumference of the track is measured. This creates a measurement or observational bias in the sample of hits along the track. Such an observational

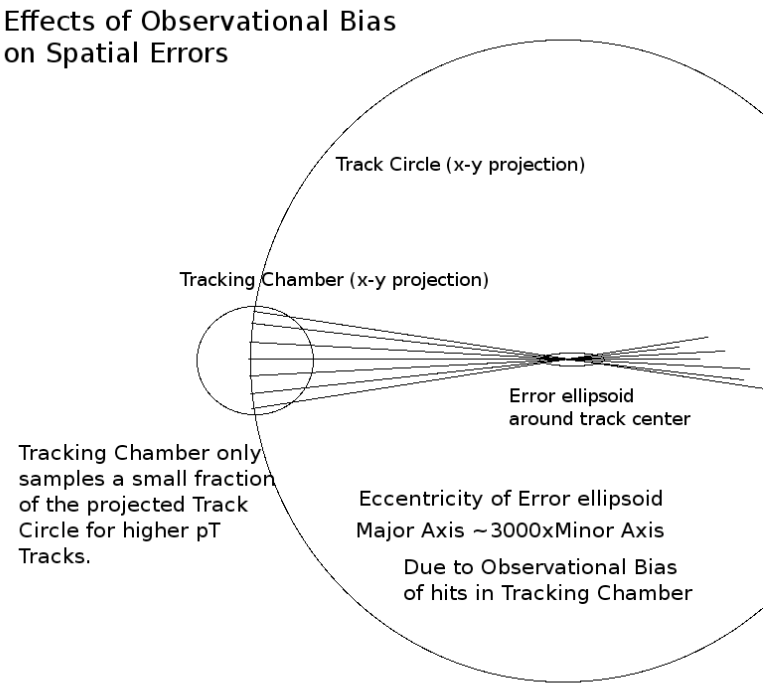

Figure 1: Relation of Error Ellipsoid of Center of Curvature and Tracking Hits.

bias creates a highly eccentric error matrix for the reconstructed x and y coordinates of the center-

of-curvature. This highly eccentric error matrix produces a spatial error matrix along the track helix which has a very small eigenvalue in the direction of the projection of the track tangent vector in the x-y plane. This property is independent of the method used to determine the track parameters precisely because the tracks we are most interested in have high p_T and, therefore, have very large diameter track circles compared to the diameter of the tracking chamber.

In order to determine if a Higgs candidate has been selected in the data-analysis of the experiment, one of the criteria applied to the 4 selected tracks is: Are these 4 tracks consistent with originating at a common position in space? Do these 4-tracks form a "vertex" in space? If the hypothetical Higgs Boson decayed into 4 charged leptons, then each lepton would follow its own helical trajectory, but all 4 leptons would converge on a common point in space corresponding to the point of decay of the Higgs Boson. Reducible 4-lepton background events, would be inconsistent with originating from the same common location. Therefore constructing a likelihood for the hypothesis that the 4 tracks originate at the same location is a very useful algorithm to separate signal events (Higgs candidates) from reducible backgrounds.

The tracks are assumed to come from a common point by hypothesis and the likelihood is obtained by first constructing the $\mathbf{K}$ matrix of Eq. (6) above with the $\mathbf{y}_i(s_i) = [x(s_i), y(s_i), z(s_i)]$ for coordinates $x(s_i)$, $y(s_i)$, and $z(s_i)$ for each track (i is numbered 1 through 4) with its respective independent variable s_1, s_2, s_3 or s_4. The $\mathbf{R}$ matrix in Eq. (6) is obtained by the "propagation of errors" method which involves Taylor expansion of the functions $x(s)$, $y(s)$ and $z(s)$ about the mean positions. The overall matrix is obtained by appending each of the contributions from each track to the form the assembly $\mathbf{y} = [\mathbf{y_1}, \mathbf{y_2}, \mathbf{y_3}, \mathbf{y_4}]$ and similarly for $\mathbf{R}$. The incidence matrix $\mathbf{X}$ is similarly constructed using $\mathbf{X_i}$= 3-by-3 unit matrices (for $i = 1, 2, 3, 4$) and stacking four such matrices one-above-the-other. Given 3-by-3 unit incidence matrices for $\mathbf{X}_i = \mathbf{I_3}$ (for the i-th track), then $\beta = \vec{r}_c$. The extraneous parameters β are the central position coordinates of the vertex which can be estimated using the method of Section 7. This relationship completes the definition of the Linear Model.

Since the extrapolated spatial error matrices have the nearly rank-2 property, it is necessary to utilize the methods described above to construct a consistent ReML for testing the hypothesis of the common spatial origin of the 4 selected leptons as well as automatically constructing the non-stochastic linear combinations which have to be constrained to zero (a la the Second Special Form

mentioned above). The constraints are determined based on comparing the size of the respective diagonal element in the Cholesky decomposition to a tunable threshold value which is optimized for signal events. When such a zero is encountered, the location on the main diagonal is noted and stored for subsequent determination of the constraint itself. The ReML likelihood and collateral constraints are determined as outlined above and the system of first and second derivatives are calculated using the Tables 1, 2, and 3 of the appendix. A linear system of equations is then formed by Taylor expansion of the objective function and a step is taken by Newton-Raphson (NR) iteration to move towards the maximum likelihood position. The coordinate functions of Eq. (7), evaluated at $s = 0$, produce the coordinates of the helix at closest approach to the beam-axis (z-axis). Since a real Higgs would be produced very close to the colliding beam axis, then s_1, s_2, s_3, s_4 should all be initialized to 0 at the beginning of the first NR step for fast convergence. Alternatively, one could perform a binary search about $s = 0$ (for each track) to determine appropriate initial values for the NR method.

After convergence, the algorithm produces a set of four parameter values s_1, s_2, s_3, s_4 which give the location along each track such that the assumptions of the model are satisfied. The χ^2 value is obtained as the square of the last element on the main diagonal of the resulting decomposed $\mathbf{K}$ matrix ($\chi^2 = -L^2_{k,k}$). This variable is only approximately χ^2 coming with degrees of freedom that are underestimated (by the number of positive pivots minus the number of negative pivots plus 1, minus 4 from estimating the position variables s_1, s_2, s_3, s_4) and we use the empirical distribution for χ^2 in applications. ReML and χ^2 are invariant with respect to the coordinates of the central position (the vertex), because the impact of the central position was removed from the likelihood. The crucial assumption of the the model is that there be a common origination point for the four tracks *somewhere*. If the χ^2 from the fit is small, the model is satisfied and the hypothesis that the 4-tracks originated from a common point is consistent with the data. This is what we expect for Higgs $\rightarrow$ 4 leptons. Reducible background events (such as $t\bar{t}, Zbb \rightarrow$ 4 leptons) where the 4-leptons do not originate from the same point will not fit the hypothesis of the model and will have large χ^2 values.

In order to implement the above algorithm a sample of events was generated using the PYTHIA (Sjöstrand, *et al.*, 2006) Event Generation program. These events were then simulated using the LDT Monte Carlo Program (Regler, *et al.*, 2007) and (Valentan, *et al.*, 2011) to determine the

detector response and to obtain fits to the track parameters and their covariance matrix. The analysis of the fitted tracks and their error matrices was done using the Rave/Vertigo datahar-vesting and vertexing environment (Waltenberger, 2011) with the above ReML algorithm inserted internally in the package.

Applying the ReML method to the case of 4 leptons from Higgs decay and also from a known major source of background ($t\bar{t} \rightarrow 4$ leptons), where the leptons can be either electrons or muons, we arrive at a comparison of signal and background shown in Fig. (2). Only events in which all four tracks had $p_T > 5$ GeV/c were used.

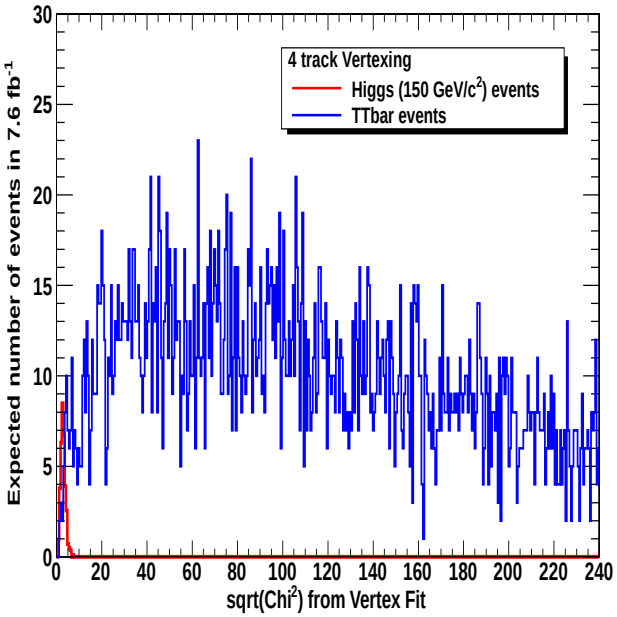

Figure 2: Red histogram: Expected number of Higgs events at $\mathcal{L} = 7.62$ femptobarn^{-1} as a function of the ReML $\sqrt{\chi^2}$. Blue histogram: Expected number of $t\bar{t}$ background events at the same luminosity.

As can be seen Fig. (2), even though the Higgs signal events have a smaller cross section and a smaller number of expected events than $t\bar{t}$ events, they are much more peaked near zero $\sqrt{\chi^2}$ for the same acceptance conditions of the detector. This results in a much higher efficiency (CDF value) for detecting Higgs as compared to retaining background events as can be seen in Fig. (3). This means that we can preferentially select Higgs candidates and reject $t\bar{t}$ background by cutting

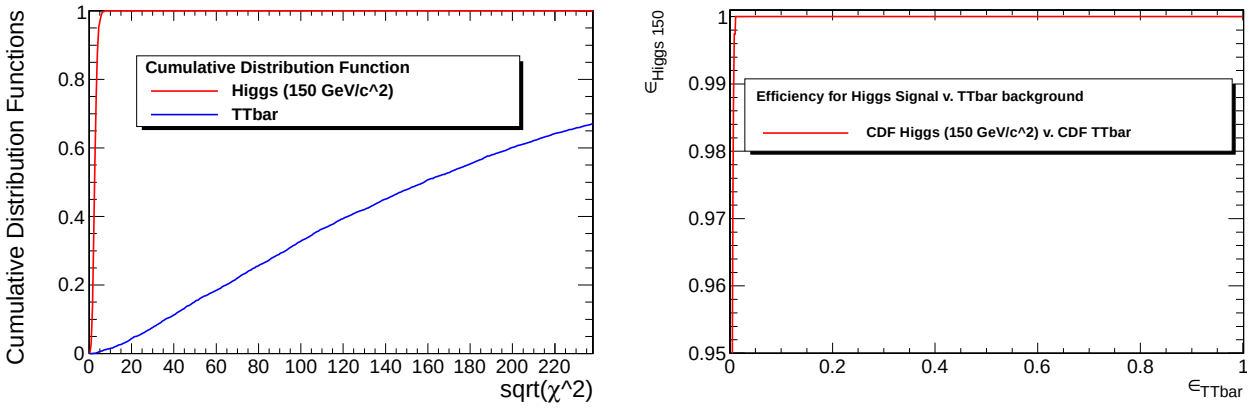

Figure 3: Left: Red histogram: Cumulative Distribution Function for Higgs Boson events as a function of the ReML $\sqrt{\chi^2}$. Blue histogram: Cumulative Distribution Function for $t\bar{t}$ background events. Right: Cumulative distributions plotted against each other at the same $\sqrt{\chi^2}$. The CDF for accepting Higgs $\rightarrow$ 4 lepton events is shown versus CDF for accepting $t\bar{t} \rightarrow$ 4 lepton events based on the χ^2 from 4-track vertexing.

on the χ^2 value.

In the case of high luminosity collisions, there can be multiple events detected every time the detector is triggered. These multiple events are recorded by the detector in the same time window and are called "pile up". The ReML method, as well as other 4-track vertexing methods, have the advantage of being robust against the effects of high pile up since it does not depend on an auxiliary determination of the best vertex (the Primary Vertex) that matches with the 4-tracks used in the above analysis. Finding the correct Primary Vertex increases with difficulty as the proton-proton collision luminosity increases. Discrimination methods, such as b-tagging (Tomalin, 2008), will be adversely affected as the number of pile up events increases. At the LHC design luminosity, approximately 200 pile up events per beam crossing are expected. The issue of finding the best Primary Vertex is irrelevant to the 4-track vertex method. If one of the 4-tracks does not match with the other 3 because it is from an un-related pile up event, then both the b-tagging algorithm and the 4-track vertex method will assign a large χ^2 to that event and it will be rejected. However, if the wrong Primary Vertex is used as a reference for the b-tagging then all 4-tracks will register a large χ^2 even though they might be matched to a common spatial position by the 4-track vertex algorithm. These events should not be discarded just because the wrong and

irrelevant Primary Vertex was used as a reference marker. If the 4-tracks have a small value of χ^2 from the 4-track vertex method, then that alone is sufficient to decide if they should be further analyzed.

9 Conclusion

This paper presents an automated matrix procedure for constructing the ReML likelihood function and also to identify non-stochastic linear combinations that represent quantities that must be constrained to zero in the event that the error matrix describing the data is singular. The method is applied to the problem of determining how close four tracks from a Higgs $\rightarrow$ 4 lepton decay approach a common position in space. This method can be used to discriminate between signal and background events at experiments at the LHC. It has an advantage of being independent of pile up which is not the case for b-tagging.

10 Acknowledgments

We are especially thankful for the assistance of Nicola De Filippis, (Politecnico and INFN Bari, Italy), who incorporated the above method in the software for the CMS Experiment. The analysis outside of the CMS environment was made possible by the work of Manfred Valentan (Institute of High Energy Physics, Austrian Academy of Sciences, Vienna, Austria) who wrote the interface to read events generated by PYTHIA and provided the LDT Monte Carlo program to simulate and reconstruct the tracks. Without the help and encouragement of Wolfgang Waltenberger (Institute of High Energy Physics, Austrian Academy of Sciences, Vienna, Austria), it would have been impossible to develop the code necessary to insert the ReML algorithm into the Rave/Vertigo analysis environment. We thank Andrey Korytov and Alexey Drozdetskiy (University of Florida, Gainesville, Florida, USA) for interesting discussions on how to compare signal and background events using the method described above on Monte Carlo samples. Thanks also go to Robert Alan Wolf (University of San Francisco Mathematics Department, San Francisco, California, USA) and Chris Freiling (California Sate University Mathematics Department, San Bernardino, California, USA) for their mathematical insights into the geometrical problems encountered in this work.

References

Bengio, Y. (2000). Continuous optimization of hyper-parameters. *Proceeding of the IEEE-INNS-ENNS International Joint Conference on Neural Networks* **1**, 305–310.

Capriotti, L., and Giles, M. (2010). Fast Correlation Greeks by Adjoint Algorithmic Differentiation. *Risk Magazine* April 2010, 79-89.

Christianson, D. B., Davies, A. J., Dixon, L. C. W., Roy, R., and Van der zee, P. (1997). Giving reverse differentiation a helping hand. *Optimization Methods and Software* **8(1)**, 53-67.

De Hoog, F.R., Anderssen, R.S. and Lukas, M.A. (2011). Differentiation of Matrix Functionals Using Triangular Factorization. *Mathematics of Computation* **80**, 1585-1600.

Frimannslund, L., and Skaug, H. J. (2006). Nested optimisation: Application to estimation of variation in annual mortality in fish populations. In Frimannslund, *On Curvature and Separability in Unconstrained Optimisation.* Ph.D. Thesis, University of Bergen, Norway (2006).

Goldberger, A. S. (1962). Best Linear Unbiased Prediction in the Generalized Linear Regression Model. *Journal of the American Statistical Association* **70**, 369–375.

Griewank, A., and Walther, A. (2008). Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation. *Society for Industrial and Applied Mathematics* (SIAM).

Harville, D. A. (1974). Bayesian Inference for Variance Components Using only Error Contrasts. *Biometrika* **61**, 383–385.

Harville, D. A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association* **72 (358)**, 320–340.

Henderson, C. R., Kempthorne, O., Searle, S. R., and Von Krosigk, C. N. (1959). Estimation of Environmental and Genetic Trends from Records Subject to Culling. *Biometrics* **15 (2)**, 192-218.

Meyer, K. (2001). Estimating genetic covariance functions assuming a parametric correlation structure for environmental effects. *Genetics Selection, Evolution* **33**, 557-585.

Meyer, K. (2007). WOMBAT – A tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). *Journal of Zhejiang University Science* **B 8(11)**, 815–821.

Meyer, K., and Kirkpatrick, M. (2005). Restricted maximum likelihood estimation of genetic principal components and smoothed covariance matrices. *Genetics Selection. Evolution* **37**, 1-30.

Meyer, K., and Smith, S. P. (1995). Restricted maximum likelihood estimation for animal models using derivatives of the likelihood. *Genetics Selection. Evolution* **28**, 23–49.

Patterson, H.D., and Thompson, R. (1971). Recovery of Inter-block Information when Block Sizes are Unequal. *Biometrika* **58**, 545–554.

Regler, M., Valentan, M., and Frühwirth, R. (2007). The LiC Detector Toy Program. Proc. 11th Vienna Conference on Instrumentation (VCI) 2007, Vienna, Austria. *Nucl. Instr. Meth.* **A 581** 553-556.

SAS Institute (2009). SAS/STAT 9.2 Users Guide (2nd Edition): Mixed Modeling. *SAS Institute.*

Schmidt, M. N. (2009). Function factorization using warped Gaussian processes. In L. Bottou and M. Littman, editors, *Proceedings of the 26th Annual International Conference on Machine Learning*, 921–928.

Siegel, H. (1965). Deferment of Computations in the Method of Least Squares. *Mathematics of Computation* **19**, 329–331.

Sjöstrand, T., Mrenna, S., and Skands, P. (2006). Pythia 6.4 Physics and Manual. (LU TP 06-13, FERMILAB-PUB-06-052-CD-T, hep-ph/0603175), 1-574.

Smith, S.P. (1995). Differentiation of the Cholesky Algorithm. *Journal of Computational and Graphical Statistics* **4 (2)**, 134-147.

Smith, S. P. (1997). Sparse matrix tools for Gaussian models on lattices. *Computational Statistics & Data Analysis* **26**, 1-15.

Smith, S. P. (2000). A Tutorial on Simplicity and Computational Differentiation for Statisticians. *U.C. Davis Physics Dept Memo*.

Smith, S. P. (2001a). Likelihood-Based Analysis of Linear State-Space Models Using the Cholesky Decomposition. *Journal of Computational and Graphical Statistics* **10 (2)**, 350–369.

Smith, S. P. (2001b). The factorability of symmetric matrices and some implications for statistical linear models. *Linear Algebra and its Applications* **335**, 63–80.

Speed, T. P. (1997). Restricted maximum likelihood (ReML). *Encylopedia of Statistical Science (C.B.Read ed.) Wiley, New York* **2**, 472–481.

Speed, T. P. (1995). ReML: A Brief Review. *Statistics Research Report*, **SRR 004-95**.

Toal, D. J. J., Forrester, A. I. J., Bressloff, N. W., Keanel, A. J., and Holden, C. (2009). An adjoint for likelihood maximization. *Proceedings of the Royal Society* **A**: Mathematics, 465, (2111), 3267-3287.

Tomalin, I. (2008). b Tagging in CMS. *Journal of Physics: Conference Series* **110**, 092033-092036. CMS Collaboration (2011). Status of b-tagging tools for 2011 data analysis *CMS PAS BTV-11-002*, 1-35.

Valentan, M., Frühwirth, R., Mitaroff, W., and Regler., M. (2011). The LiC Detector Toy fast simulation program. The 20th Anniversary International Workshop on Vertex Detectors - VERTEX 2011, Rust, Lake Neusiedl, Austria. *To be published in Proceedings of Science*

Waltenberger, W. (2011). Rave – A Detector-Independent Toolkit to Reconstruct Vertices. *IEEE Transactions on Nuclear Science* **58 (2)**, 434-444.

Wikipedia (2011). http://en.wikipedia.org/wiki/Restricted_maximum_likelihood

11 Appendix: Pseudocode for Differentiation of Cholesky Decomposition

This Section includes corrections (shown in red) to the pseudocode which was given in (Smith, 2001a). Construction of the Cholesky decomposition for an indefinite matrix $\mathbf{K}_{N\times N}$ is summarized in the following table

Initializations:

$\mathbf{L}$ is lower triangular and $L_{ij} \leftarrow ij$-th element of $\mathbf{K}_{N\times N}(x_1, x_2, ..., x_p)$.

$\ominus_k$ = "+" if k-th diagonal is part of non-negative submatrix, "−" otherwise.

$\pm_k = -\ominus_k$.

Algorithm:

For $k = 1, ..., N$ do

if $|L_{kk}| \approx$ zero, check to see if remaining $L_{jk} \approx$ zero, $j = k+1, ..., N$ and skip k-th pivot.

Otherwise, $L_{kk} \leftarrow \text{sqrt}[\ominus_k L_{kk}]$ and do:

$L_{jk} \leftarrow \ominus_k L_{jk}/L_{kk}$, for $j = k+1, ..., N$,

$L_{ij} \leftarrow L_{ij} \pm_k (L_{jk} \times L_{ik})$, for $j = k+1, ..., N$ & $i = j, ..., N$.

end k

Table 1. Pseudocode for Cholesky Decomposition with Possible Negative Diagonals.

First derivatives, $\partial F(\mathbf{L})/\partial x_v$, are summarized in the following table

Initializations:

$\mathbf{L}$ is provided (see Table 1).

$\mathbf{F}_{N\times N}$ is a work space with elements F_{ij}, defined respectively for $i \geq j$.

$\ominus_k$ = "+" if k-th diagonal is part of non-negative submatrix, "−" otherwise.

$\pm_k = -\ominus_k$.

$F_{ij} \leftarrow \partial F(\mathbf{L})/\partial L_{ij}$

Algorithm:

(a) $\mathbf{F} \leftarrow T(\mathbf{F})$ by following operations:

For $k = N, ..., 1$ (N.B. Decreasing Order) do

if $|L_{kk}| >$ zero, then do:

$F_{ik} \leftarrow F_{ik} \pm_k (F_{ij} \times L_{jk})$, $F_{jk} \leftarrow F_{jk} \pm_k (F_{ij} \times L_{ik})$, for $j = k+1, ..., N$ & $i = j, ..., N$

$F_{jk} \leftarrow \ominus_k F_{jk}/L_{kk}$, $F_{kk} \leftarrow F_{kk} \pm_k (F_{jk} \times L_{jk})$, for $j = k+1, ..., N$

$F_{kk} \leftarrow \ominus_k (1/2) F_{kk}/L_{kk}$

end k

(b) $\partial F(\mathbf{L})/\partial x_v = \sum_{i\geq j} F_{ij} \times \partial K_{ij}/\partial x_v, \quad v = 1, 2, ...p.$

Table 2. Pseudocode for Backward Differentiation of $F(\mathbf{L})$.

Second derivatives, $\partial^2 F(\mathbf{L})/\partial x_v \partial x_u$, are summarized in the following table

Initializations:

$\mathbf{L}$ is provided (see Table 1).

$\ominus_k$ = "+" if k-th diagonal is part of non-negative submatrix, "−" otherwise.

$\pm_k = -\ominus_k$.

$\mathbf{F} \leftarrow T(\partial F(\mathbf{L})/\partial L_{ij})$ provided (see Table 2).

$\mathbf{S}_{N\times N}$ and $\mathbf{Q}_{N\times N}$ are work spaces with elements S_{ij}, Q_{ij}, defined respectively for $i \geq j$.

$Q_{ij} \leftarrow \partial K_{ij}/\partial x_v, \quad i \geq j, \quad v \in \{1, 2, ..., p\}$.

Algorithm:

(a) For $k = 1, ..., N$ do forward sweep

if $|L_{kk}| >$ zero, then do:

$Q_{kk} \leftarrow \ominus_k(1/2) \times Q_{kk}/L_{kk}$,

$S_{kk} \leftarrow \pm_k 2 \times Q_{kk}F_{kk}$,

$Q_{jk} \leftarrow [\ominus_k Q_{jk} - Q_{kk}L_{jk}]/L_{kk}$,

$S_{jk} \leftarrow \pm_k Q_{kk} \times F_{jk}$,

$S_{kk} \leftarrow S_{kk} \pm_k (Q_{jk} \times F_{jk})$, for $j = k + 1, ..., N$

$Q_{ij} \leftarrow Q_{ij} \pm_k (Q_{ik} \times L_{jk}) \pm_k (L_{ik}Q_{jk})$,

$S_{ik} \leftarrow S_{ik} \pm_k (F_{ij} \times Q_{jk}), S_{jk} \leftarrow S_{jk} \pm_k (F_{ij} \times Q_{ik})$, for $j = k + 1, ..., N$ & $i = j, ..., N$

(b) for $i \geq j, S_{ij} \leftarrow S_{ij} + \sum_{m\geq n} Q_{mn} \times \partial^2 F(\mathbf{L})/\partial L_{mn}\partial L_{ij}$.

(c) reverse sweep, $\mathbf{S} \leftarrow T(\mathbf{S})$ (see Table 2).

(d) $\partial^2 F(\mathbf{L})/\partial x_v \partial x_u = \sum_{i\geq j} S_{ij} \times \partial K_{ij}/\partial x_u + \sum_{i\geq j} F_{ij} \times \partial^2 K_{ij}/\partial x_v \partial x_u, u = 1, 2, ..., p.$

Table 3. Pseudocode for Backward Differentiation Applied Twice to $F(\mathbf{L})$.