On The Direct Maximization of Quadratic Weighted Kappa

David Vaughn DVAUGHN@MEASINC.COM
Derek Justice DJUSTICE@MEASINC.COM

Abstract

In recent years, quadratic weighted kappa has been growing in popularity in the machine learning community as an evaluation metric in domains where the target labels to be predicted are drawn from integer ratings, usually obtained from human experts. For example, it was the metric of choice in several recent, high profile machine learning contests hosted on Kaggle : www.kaggle.com/c/asap-aes , www.kaggle.com/c/asap-sas , www.kaggle.com/c/diabetic-retinopathy-detection .

Yet, little is understood about the nature of this metric, its underlying mathematical properties, where it fits among other common evaluation metrics such as mean squared error (MSE) and correlation, or if it can be optimized analytically, and if so, how. Much of this is due to the cumbersome way that this metric is commonly defined.

In this paper we first derive an equivalent but much simpler, and more useful, definition for quadratic weighted kappa, and then employ this alternate form to address the above issues.

1. Preliminaries

Although first developed in the statistical community as a measure of *inter-rater agreement*, κ has more recently become a popular performance metric in supervised machine learning, specifically in situations where the target (dependent) variable $\mathbf{y}$ is a discrete, interval variable (usually drawn from non-negative integers) such as is common in most human rating scales (e.g."on a scale from 1 to 10").

This differs from the ordinal regression setting, where there only exists an ordering over labels, but no intrinsic or constant length interval between them. Some have argued that the use of quadratic weighted kappa as a metric in the domain of human ratings imposes the erroneous assumption of "equal intervals" where there should be none (how this assumption is expressed in the metric itself will be made clear in the following section). For example, when rating student essays on a scale from 1 to 5, the difference between a 1 and a 2 may not be equal to the difference between a 4 and a 5. While this may or may not be true in certain cases, we will not be concerned with that here.

1.1. Standard Definition

Quadratic weighted kappa, which we write κ to distinguish from linear weighted kappa, was originally developed as a measure of *inter-rater agreement*. In this scenario, there are two raters, $\mathcal{A}$ and $\mathcal{B}$, each associated with a vector of n integer ratings $\mathbf{a}, \mathbf{b} \in \mathbb{L}^{n\times 1}$ where $\mathbb{L} = \{1, 2, \cdots, \ell\}$ is a finite set of ℓ possible values. We seek to quantify the level of agreement between $\mathbf{a}$ and $\mathbf{b}$. In order to compute $\kappa(\mathbf{a}, \mathbf{b})$, it is customary to start by computing frequency tables. The *observed* confusion matrix $\mathbf{U} = (u_{i,j}) \in \mathbb{N}^{\ell\times\ell}$ is first computed as:

$$u_{i,j} = \sum_{k=1}^{n} \mathbb{I}(a_k = i) \cdot \mathbb{I}(b_k = j)$$

Next, the *expected* confusion matrix $\mathbf{V} = (v_{i,j}) \in \mathbb{R}^{\ell\times\ell}$ is computed by assuming there is no correlation between raters. Under this assumption, we can simply compute $\mathbf{V}$ as the outer product between the two rater's observed marginal distributions, normalized so that $\mathbf{V}$ has the same sum as $\mathbf{U}$ $\left(i.e.\ \sum_{i,j=1}^{\ell} v_{i,j} = \sum_{i,j=1}^{\ell} u_{i,j}\right)$:

$$\mathbf{V} = \frac{(\mathbf{U}\cdot\mathbf{e}) \otimes (\mathbf{U}^{\top}\cdot\mathbf{e})}{n}$$

where $\mathbf{e}$ denotes the all ones vector. Finally, a matrix of weights $\mathbf{W} = (w_{i,j}) \in \mathbb{R}^{\ell\times\ell}$ is defined as:

$$w_{i,j} = \frac{(i-j)^2}{(\ell-1)^2}$$

Given $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{W}$, it is customary to define κ as:

$$\kappa(\mathbf{a}, \mathbf{b}) = 1 - \frac{\sum_{i,j=1}^{\ell} u_{i,j}\cdot w_{i,j}}{\sum_{i,j=1}^{\ell} v_{i,j}\cdot w_{i,j}} = 1 - \frac{\langle \mathbf{U}, \mathbf{W}\rangle_F}{\langle \mathbf{V}, \mathbf{W}\rangle_F} \quad (1)$$

where $\langle\cdot,\cdot\rangle_F$ denotes the Frobenius inner product.

1.2. Alternate Form

Assume a data matrix $\mathbf{X} = [\mathbf{x}_1^\top, \cdots, \mathbf{x}_n^\top]^\top \in \mathbb{R}^{n \times d}$ of n points in d dimensional space, and a vector $\mathbf{y} \in \mathbb{L}^{n \times 1}$ of n integer labels where $\mathbb{L} = \{1, 2, \cdots, \ell\}$ is a finite set of possible labels. We assume there is some true functional relationship $\boldsymbol{y}_i = f(\boldsymbol{x}_i)$ between each point $\boldsymbol{x}_i$ and it's label $\boldsymbol{y}_i$. The goal is to find a function $\hat{\boldsymbol{y}}_i = \hat{f}(\boldsymbol{x}_i)$ which approximates the true function as closely as possible. Using this notation, we can begin to rewrite the standard definition of κ by first noting that the numerator simply represents a standard sum of squared errors:

$$\langle \mathbf{U}, \mathbf{W} \rangle_{_F} = \sum_{k=1}^{n} \frac{(\boldsymbol{y}_k - \hat{f}(\boldsymbol{x}_k))^2}{(\ell - 1)^2} = \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{(\ell - 1)^2}$$

Next, we note that the denominator can be similarly rewritten in terms of $\mathbf{y}$ and $\hat{\mathbf{y}}$:

$$\langle \mathbf{V}, \mathbf{W} \rangle_{_F} = \frac{\|\mathbf{y}\|^2 - \frac{2}{n}(\mathbf{y}^\top \mathbf{e})(\hat{\mathbf{y}}^\top \mathbf{e}) + \|\hat{\mathbf{y}}\|^2}{(\ell - 1)^2}$$

By combining these expressions we can derive a simplified expression for κ purely in terms of $\mathbf{y}$ and $\hat{\mathbf{y}}$, without the need for contingency tables:

$$\begin{aligned}
\kappa(\mathbf{y}, \hat{\mathbf{y}}) &= 1 - \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{\|\mathbf{y}\|^2 - \frac{2}{n}(\mathbf{y}^\top \mathbf{e})(\hat{\mathbf{y}}^\top \mathbf{e}) + \|\hat{\mathbf{y}}\|^2} \\
&= 1 - \frac{\|\mathbf{y}\|^2 - 2\mathbf{y}^\top \hat{\mathbf{y}} + \|\hat{\mathbf{y}}\|^2}{\|\mathbf{y}\|^2 - \frac{2}{n}(\mathbf{y}^\top \mathbf{e})(\hat{\mathbf{y}}^\top \mathbf{e}) + \|\hat{\mathbf{y}}\|^2} \\
&= \frac{2[\mathbf{y} - \mathbf{e}(\frac{\mathbf{y}^\top \mathbf{e}}{n})]^\top \hat{\mathbf{y}}}{\|\mathbf{y}\|^2 - \frac{2}{n}(\mathbf{y}^\top \mathbf{e})(\hat{\mathbf{y}}^\top \mathbf{e}) + \|\hat{\mathbf{y}}\|^2}
\end{aligned}$$

If we drop the integer constraint on labels then we can assume that $\mathbf{y}$ has been centered, which allows further simplification:

$$\kappa(\mathbf{y}, \hat{\mathbf{y}}) = \frac{2\langle \mathbf{y} \cdot \hat{\mathbf{y}} \rangle}{\|\mathbf{y}\|^2 + \|\hat{\mathbf{y}}\|^2} \tag{2}$$

Throughout the rest of this paper we will assume (without loss of generality) that $\mathbf{y}$ is centered.

2. Linear Model

We consider a simple linear model for $\mathbf{y} = f(\mathbf{X})$:

$$f(\mathbf{X}) = \mathbf{X}\boldsymbol{\alpha}$$

and so we seek the function parameter $\hat{\boldsymbol{\alpha}}$ that maximizes $\kappa(\boldsymbol{\alpha})$, where

$$\kappa(\boldsymbol{\alpha}) = \frac{2\langle \mathbf{y} \cdot \mathbf{X}\boldsymbol{\alpha} \rangle}{\|\mathbf{y}\|^2 + \|\mathbf{X}\boldsymbol{\alpha}\|^2}$$

We will rely on the fact that $\kappa(\boldsymbol{\alpha})$ is *strictly quasiconcave* (see appendix for proof), since it is known that any local maximum of a *strictly quasiconcave* function f is also a global maximum (Avriel, 2003). And so we can apply the method of Lagrange multipliers to find a local maximum, and conclude that this is the global maximum as well. To begin, we can use the variable substitution $\gamma = \|\mathbf{y}\|^2 + \|\mathbf{X}\boldsymbol{\alpha}\|^2$ to write this as a constrained optimization problem:

$$\begin{aligned}
\max_{\boldsymbol{\alpha}, \gamma} \quad & \frac{2}{\gamma}\langle \mathbf{y} \cdot \mathbf{X}\boldsymbol{\alpha} \rangle \\
\text{subject to} \quad & \gamma = \|\mathbf{y}\|^2 + \|\mathbf{X}\boldsymbol{\alpha}\|^2
\end{aligned}$$

From here we can define the Lagrangian:

$$\begin{aligned}
\mathbf{L}(\boldsymbol{\alpha}, \gamma, \lambda) &= \frac{2}{\gamma}\langle \mathbf{y} \cdot \mathbf{X}\boldsymbol{\alpha} \rangle + \lambda(\gamma - \|\mathbf{y}\|^2 - \|\mathbf{X}\boldsymbol{\alpha}\|^2) \\
&= \frac{2}{\gamma}\mathbf{y}^\top \mathbf{X}\boldsymbol{\alpha} + \lambda(\gamma - \mathbf{y}^\top \mathbf{y} - \boldsymbol{\alpha}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\alpha})
\end{aligned}$$

We solve by first differentiating w.r.t. γ and setting equal to zero:

$$\begin{aligned}
\left.\frac{\partial \mathbf{L}}{\partial \gamma}\right|_{\hat{\boldsymbol{\alpha}}, \hat{\gamma}, \hat{\lambda}} &= -\frac{2}{\hat{\gamma}^2}\mathbf{y}^\top \mathbf{X}\hat{\boldsymbol{\alpha}} + \hat{\lambda} = 0 \\
\hat{\lambda} &= \frac{2}{\hat{\gamma}^2}\mathbf{y}^\top \mathbf{X}\hat{\boldsymbol{\alpha}}
\end{aligned} \tag{3}$$

Then w.r.t. $\boldsymbol{\alpha}$:

$$\left.\frac{\partial \mathbf{L}}{\partial \boldsymbol{\alpha}}\right|_{\hat{\boldsymbol{\alpha}}, \hat{\gamma}, \hat{\lambda}} = \frac{2}{\hat{\gamma}}\mathbf{X}^\top \mathbf{y} - 2\hat{\lambda}\mathbf{X}^\top \mathbf{X}\hat{\boldsymbol{\alpha}} = 0$$

Substituting in the previous expression (3) for $\hat{\lambda}$ yields:

$$\begin{aligned}
\frac{\mathbf{X}^\top \mathbf{y}}{\hat{\gamma}} - \frac{2\mathbf{y}^\top \mathbf{X}\hat{\boldsymbol{\alpha}}}{\hat{\gamma}^2}\mathbf{X}^\top \mathbf{X}\hat{\boldsymbol{\alpha}} &= 0 \\
\frac{2\mathbf{y}^\top \mathbf{X}\hat{\boldsymbol{\alpha}}}{\hat{\gamma}}\mathbf{X}^\top \mathbf{X}\hat{\boldsymbol{\alpha}} &= \mathbf{X}^\top \mathbf{y}
\end{aligned}$$

Now substituting back in for $\hat{\gamma}$ gives a surprising result:

$$\begin{aligned}
\frac{2\mathbf{y}^\top \mathbf{X}\hat{\boldsymbol{\alpha}}}{\|\mathbf{y}\|^2 + \|\mathbf{X}\hat{\boldsymbol{\alpha}}\|^2}\mathbf{X}^\top \mathbf{X}\hat{\boldsymbol{\alpha}} &= \mathbf{X}^\top \mathbf{y} \\
\hat{\kappa}\mathbf{X}^\top \mathbf{X}\hat{\boldsymbol{\alpha}} &= \mathbf{X}^\top \mathbf{y} \\
\hat{\boldsymbol{\alpha}} &= \frac{1}{\hat{\kappa}}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}
\end{aligned} \tag{4}$$

where we have simplified by recognizing the expression for $\hat{\kappa} = \kappa(\hat{\boldsymbol{\alpha}})$. In addition, we recognize the "least-squares" solution $\hat{\boldsymbol{\alpha}}_{\ell s} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$, the well studied minimizer of the quantity $\|\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}\|$, *a.k.a.* the norm of the residual. This means that $\hat{\boldsymbol{\alpha}}$ (which we will write $\hat{\boldsymbol{\alpha}}_\kappa$ when necessary to avoid confusion) and the least-squares solution $\hat{\boldsymbol{\alpha}}_{\ell s}$ are related via $\hat{\kappa}$ by a surprisingly simple formula:

$$\hat{\boldsymbol{\alpha}}_\kappa = \frac{1}{\hat{\kappa}}\hat{\boldsymbol{\alpha}}_{\ell s} \tag{5}$$

In order to obtain a closed form solution for $\hat{\boldsymbol{\alpha}}_\kappa$ we must solve for $\hat{\kappa}$ solely in terms of $\mathbf{X}$ and $\mathbf{y}$. We employ another useful tool from least-squares regression, the "hat" matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$. The hat matrix allows the least-squares labels $\hat{\mathbf{y}}_{\ell s}$ to be expressed solely in terms of the original labels $\mathbf{y}$:

$$\hat{\mathbf{y}}_{\ell s} = \mathbf{X}\hat{\boldsymbol{\alpha}}_{\ell s} = \mathbf{H}\mathbf{y} \tag{6}$$

Using the well-known fact that the hat matrix is both symmetric and idempotent ($\mathbf{H}^\top\mathbf{H} = \mathbf{H}^2 = \mathbf{H}$), we can establish the following identity:

$$\begin{aligned}\|\hat{\mathbf{y}}_{\ell s}\|^2 &= (\mathbf{H}\mathbf{y})^\top(\mathbf{H}\mathbf{y}) \\ &= \mathbf{y}^\top\mathbf{H}^\top\mathbf{H}\mathbf{y} \\ &= \mathbf{y}^\top\mathbf{H}\mathbf{y} \\ &= \mathbf{y}^\top\mathbf{X}\hat{\boldsymbol{\alpha}}_{\ell s} = \langle\mathbf{y}\cdot\mathbf{X}\hat{\boldsymbol{\alpha}}_{\ell s}\rangle\end{aligned} \tag{7}$$

Now we can take the derived expression for $\hat{\boldsymbol{\alpha}}_\kappa$ from (5) and substitute it back into the original expression for κ and apply the above identity (7) to solve for $\hat{\kappa}$:

$$\begin{aligned}\hat{\kappa} = \kappa(\hat{\boldsymbol{\alpha}}_\kappa) &= \frac{2\langle\mathbf{y}\cdot\mathbf{X}\hat{\boldsymbol{\alpha}}_\kappa\rangle}{\|\mathbf{y}\|^2 + \|\mathbf{X}\hat{\boldsymbol{\alpha}}_\kappa\|^2} \\ &= \frac{2\langle\mathbf{y}\cdot\frac{1}{\hat{\kappa}}\mathbf{X}\hat{\boldsymbol{\alpha}}_{\ell s}\rangle}{\|\mathbf{y}\|^2 + \|\frac{1}{\hat{\kappa}}\mathbf{X}\hat{\boldsymbol{\alpha}}_{\ell s}\|^2} \\ &= \frac{\frac{2}{\hat{\kappa}}\langle\mathbf{y}\cdot\mathbf{X}\hat{\boldsymbol{\alpha}}_{\ell s}\rangle}{\|\mathbf{y}\|^2 + \frac{1}{\hat{\kappa}^2}\|\mathbf{X}\hat{\boldsymbol{\alpha}}_{\ell s}\|^2} \\ &= \frac{2\hat{\kappa}\|\hat{\mathbf{y}}_{\ell s}\|^2}{\hat{\kappa}^2\|\mathbf{y}\|^2 + \|\hat{\mathbf{y}}_{\ell s}\|^2} \\ \hat{\kappa}^2\|\mathbf{y}\|^2 + \|\hat{\mathbf{y}}_{\ell s}\|^2 &= 2\|\hat{\mathbf{y}}_{\ell s}\|^2 \\ \hat{\kappa} &= \frac{\|\hat{\mathbf{y}}_{\ell s}\|}{\|\mathbf{y}\|}\end{aligned} \tag{8}$$

Finally we can express $\hat{\boldsymbol{\alpha}}$ solely in terms of $\mathbf{X}$ and $\mathbf{y}$:

$$\begin{aligned}\hat{\boldsymbol{\alpha}}_\kappa = \frac{1}{\hat{\kappa}}\hat{\boldsymbol{\alpha}}_{\ell s} &= \frac{\|\mathbf{y}\|}{\|\hat{\mathbf{y}}_{\ell s}\|}\hat{\boldsymbol{\alpha}}_{\ell s} = \frac{\|\mathbf{y}\|}{\|\hat{\mathbf{y}}_{\ell s}\|}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \\ &= \left(\sqrt{\frac{\mathbf{y}^\top\mathbf{y}}{\mathbf{y}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}}}\right)(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \\ &= \left(\sqrt{\frac{\mathbf{y}^\top\mathbf{y}}{\mathbf{y}^\top\mathbf{H}\mathbf{y}}}\right)(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}\end{aligned} \tag{9}$$

2.1. Regularization

It is often useful to include a regularization term in the optimization to account for cases where $\mathbf{X}^\top\mathbf{X}$ is singular or ill-conditioned, which could lead to a numerically unstable (or non-existent) solution. In addition, it provides a convenient mechanism for manually tuning the bias/variance trade-off. It is easy to accommodate a norm penalty term within the current framework, yielding a slightly modified optimization problem:

$$\begin{aligned}\max_{\boldsymbol{\alpha},\gamma} \quad & \frac{2}{\gamma}\langle\mathbf{y}\cdot\mathbf{X}\boldsymbol{\alpha}\rangle - \mu\|\boldsymbol{\alpha}\|^2 \\ \text{subject to} \quad & \gamma = \|\mathbf{y}\|^2 + \|\mathbf{X}\boldsymbol{\alpha}\|^2\end{aligned}$$

where μ controls the magnitude of the penalty, and is usually set experimentally (e.g. with cross-validation). This leads to a similarly modified Lagrangian:

$$\mathbf{L}(\boldsymbol{\alpha},\gamma,\lambda) = \frac{2}{\gamma}\mathbf{y}^\top\mathbf{X}\boldsymbol{\alpha} - \mu\boldsymbol{\alpha}^\top\boldsymbol{\alpha} + \lambda(\gamma - \mathbf{y}^\top\mathbf{y} - \boldsymbol{\alpha}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\alpha})$$

Proceeding in the same manner as the last section, differentiating w.r.t. γ and setting equal to zero yields the same result (3) for $\hat{\lambda}$. Differentiating w.r.t. $\boldsymbol{\alpha}$ yields a slightly different result:

$$\left.\frac{\partial\mathbf{L}}{\partial\boldsymbol{\alpha}}\right|_{\hat{\boldsymbol{\alpha}},\hat{\gamma},\hat{\lambda}} = \frac{2}{\hat{\gamma}}\mathbf{X}^\top\mathbf{y} - 2\hat{\lambda}\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\alpha}} - 2\mu\hat{\boldsymbol{\alpha}} = 0$$

Again, substituting in for $\hat{\lambda}$ and manipulating the result, we end up with:

$$\begin{aligned}\hat{\kappa}\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\alpha}} + \mu\hat{\gamma}\hat{\boldsymbol{\alpha}} &= \mathbf{X}^\top\mathbf{y} \\ \hat{\boldsymbol{\alpha}} &= \frac{1}{\hat{\kappa}}(\mathbf{X}^\top\mathbf{X} + \frac{\mu\hat{\gamma}}{\hat{\kappa}}\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}\end{aligned}$$

In order to obtain a closed form solution for $\hat{\boldsymbol{\alpha}}$ all that is important to recognize is that we once again know the optimal solution to within a factor $\varphi = \frac{1}{\hat{\kappa}}$, just as in the regularized case. The only difference is that now, instead of $\hat{\boldsymbol{\alpha}}$ just being within a scalar multiple of the least-squares solution,

$\hat{\boldsymbol{\alpha}}$ is now within a scalar multiple of the *regularized* least squares solution, i.e. the *ridge regression* solution $\hat{\boldsymbol{\alpha}}_{rr}$. As μ is just a regularization tuning parameter, it can simply absorb the $\hat{\kappa}$ and $\hat{\gamma}$. We make the substitution $\sigma = \frac{\mu\hat{\gamma}}{\hat{\kappa}}$ to make this more clear:

$$\hat{\boldsymbol{\alpha}}_{\kappa} = \frac{1}{\hat{\kappa}}(\mathbf{X}^{\top}\mathbf{X} + \sigma\mathbf{I})^{-1}\mathbf{X}^{\top}\mathbf{y} = \frac{1}{\hat{\kappa}}\hat{\boldsymbol{\alpha}}_{rr} \tag{10}$$

Just as in the previous case, we can take the derived expression for $\hat{\boldsymbol{\alpha}}_{\kappa}$ from (10) and substitute it back into the original expression for κ to solve for $\hat{\kappa}$:

$$\begin{aligned}
\hat{\kappa} = \kappa(\hat{\boldsymbol{\alpha}}_{\kappa}) &= \frac{2\langle \mathbf{y} \cdot \mathbf{X}\hat{\boldsymbol{\alpha}}_{\kappa}\rangle}{\|\mathbf{y}\|^2 + \|\mathbf{X}\hat{\boldsymbol{\alpha}}_{\kappa}\|^2} \\
&= \frac{2\langle \mathbf{y} \cdot \frac{1}{\hat{\kappa}}\mathbf{X}\hat{\boldsymbol{\alpha}}_{rr}\rangle}{\|\mathbf{y}\|^2 + \|\frac{1}{\hat{\kappa}}\mathbf{X}\hat{\boldsymbol{\alpha}}_{rr}\|^2} \\
&= \frac{\frac{2}{\hat{\kappa}}\langle \mathbf{y} \cdot \mathbf{X}\hat{\boldsymbol{\alpha}}_{rr}\rangle}{\|\mathbf{y}\|^2 + \frac{1}{\hat{\kappa}^2}\|\mathbf{X}\hat{\boldsymbol{\alpha}}_{rr}\|^2} \\
\hat{\kappa}^2\|\mathbf{y}\|^2 + \|\mathbf{X}\hat{\boldsymbol{\alpha}}_{rr}\|^2 &= 2\langle \mathbf{y} \cdot \mathbf{X}\hat{\boldsymbol{\alpha}}_{rr}\rangle \\
\hat{\kappa} &= \frac{\sqrt{2\langle \mathbf{y} \cdot \mathbf{X}\hat{\boldsymbol{\alpha}}_{rr}\rangle - \|\mathbf{X}\hat{\boldsymbol{\alpha}}_{rr}\|^2}}{\|\mathbf{y}\|}
\end{aligned} \tag{11}$$

In this case the result does not simplify as nicely as in the non-regularized case, due to the fact that $\mathbf{H}_{\sigma}$ (where $\mathbf{H}_{\sigma} = \mathbf{X}(\mathbf{X}^{\top}\mathbf{X} + \sigma\mathbf{I})^{-1}\mathbf{X}^{\top}$ denotes the regularized variant of the "hat" matrix) is no longer idempotent. Still, we can use $\mathbf{H}_{\sigma}$ to write a somewhat simplified expression for $\hat{\kappa}$:

$$\hat{\kappa} = \sqrt{\frac{2\mathbf{y}^{\top}\mathbf{H}_{\sigma}\mathbf{y} - \mathbf{y}^{\top}\mathbf{H}_{\sigma}^2\mathbf{y}}{\mathbf{y}^{\top}\mathbf{y}}} \tag{12}$$

Finally we can express the regularized solution $\hat{\boldsymbol{\alpha}}_{\kappa}$ in closed form:

$$\begin{aligned}
\hat{\boldsymbol{\alpha}}_{\kappa} &= \frac{1}{\hat{\kappa}}\hat{\boldsymbol{\alpha}}_{rr} \\
&= \left(\sqrt{\frac{\mathbf{y}^{\top}\mathbf{y}}{2\mathbf{y}^{\top}\mathbf{H}_{\sigma}\mathbf{y} - \mathbf{y}^{\top}\mathbf{H}_{\sigma}^2\mathbf{y}}}\right)(\mathbf{X}^{\top}\mathbf{X} + \sigma\mathbf{I})^{-1}\mathbf{X}^{\top}\mathbf{y}
\end{aligned} \tag{13}$$

3. Link to Correlation

There is an interesting link between quadratic weighted kappa and correlation. There are many (equivalent) definitions of correlation, but the simplest is just the normalized dot product between two vectors $\mathbf{u}, \mathbf{v}$:

$$\rho(\mathbf{u}, \mathbf{v}) = \frac{\langle \mathbf{u} \cdot \mathbf{v}\rangle}{\|\mathbf{u}\|\|\mathbf{v}\|}$$

Now returning to the (non-regularized) κ-optimal solution labels $\hat{\mathbf{y}}_{\kappa} = \mathbf{X}\hat{\boldsymbol{\alpha}}_{\kappa}$, we can expand the expression for $\rho(\mathbf{y}, \hat{\mathbf{y}}_{\kappa})$ and apply (7) and (8) to simplify:

$$\begin{aligned}
\rho(\mathbf{y}, \hat{\mathbf{y}}_{\kappa}) &= \frac{\langle \mathbf{y} \cdot \hat{\mathbf{y}}_{\kappa}\rangle}{\|\mathbf{y}\|\|\hat{\mathbf{y}}_{\kappa}|} \\
&= \frac{\langle \mathbf{y} \cdot \mathbf{X}\hat{\boldsymbol{\alpha}}_{\kappa}\rangle}{\|\mathbf{y}\|\|\mathbf{X}\hat{\boldsymbol{\alpha}}_{\kappa}|} \\
&= \frac{\frac{1}{\hat{\kappa}}\langle \mathbf{y} \cdot \mathbf{X}\hat{\boldsymbol{\alpha}}_{\ell s}\rangle}{\|\mathbf{y}\|\|\frac{1}{\hat{\kappa}}\mathbf{X}\hat{\boldsymbol{\alpha}}_{\ell s}\|} \\
&= \frac{\frac{1}{\hat{\kappa}}\|\hat{\mathbf{y}}_{\ell s}\|^2}{\frac{1}{\hat{\kappa}}\|\mathbf{y}\|\|\hat{\mathbf{y}}_{\ell s}\|} \\
&= \frac{\|\hat{\mathbf{y}}_{\ell s}\|}{\|\mathbf{y}\|} = \kappa(\mathbf{y}, \hat{\mathbf{y}}_{\kappa}) = \hat{\kappa} \quad .
\end{aligned} \tag{14}$$

And so while correlation is, in general, not equivalent to quadratic weighted kappa, they do coincide precisely at the point where kappa is maximized.

Appendix: Proof of Strict Quasiconcavity

We wish to show that the function

$$\kappa(\boldsymbol{\alpha}) = \frac{2\langle \mathbf{y} \cdot \mathbf{X}\boldsymbol{\alpha}\rangle}{\|\mathbf{y}\|^2 + \|\mathbf{X}\boldsymbol{\alpha}\|^2}$$

is *strictly quasiconcave*. A function f is said to be *strictly quasiconcave* if, for any two points $\mathbf{u}, \mathbf{v}$ and any positive weight $\theta < 1$, the following is true (Avriel, 2003):

$$f(\theta\mathbf{u} + (1-\theta)\mathbf{v}) > min[f(\mathbf{u}), f(\mathbf{v})] \tag{A-1}$$

Let us assume, without loss of generality, that $min[\kappa(\mathbf{u}), \kappa(\mathbf{v})] = \kappa(\mathbf{u})$, which leads us to write:

$$\begin{aligned}
\kappa(\mathbf{v}) &\geq \kappa(\mathbf{u}) \\
\frac{2\langle \mathbf{y} \cdot \mathbf{X}\mathbf{v}\rangle}{\|\mathbf{y}\|^2 + \|\mathbf{X}\mathbf{v}\|^2} &\geq \frac{2\langle \mathbf{y} \cdot \mathbf{X}\mathbf{u}\rangle}{\|\mathbf{y}\|^2 + \|\mathbf{X}\mathbf{u}\|^2} \\
\mathbf{y}^\top\mathbf{X}\mathbf{v} &\geq \mathbf{y}^\top\mathbf{X}\mathbf{u}\left(\frac{\|\mathbf{y}\|^2 + \|\mathbf{X}\mathbf{v}\|^2}{\|\mathbf{y}\|^2 + \|\mathbf{X}\mathbf{u}\|^2}\right)
\end{aligned} \tag{A-2}$$

Returning to the the definition of *strictly quasiconcave* and expanding the expression for $\kappa(\theta\mathbf{u} + (1-\theta)\mathbf{v})$ gives us:

$$\kappa(\theta\mathbf{u} + (1-\theta)\mathbf{v}) = \frac{2\theta\mathbf{y}^\top\mathbf{X}\mathbf{u} + 2(1-\theta)\mathbf{y}^\top\mathbf{X}\mathbf{v}}{\|\mathbf{y}\|^2 + \|\theta\mathbf{X}\mathbf{u} + (1-\theta)\mathbf{X}\mathbf{v}\|^2}$$

Now, we know that the function $h(\mathbf{z}) = \|\mathbf{X}\mathbf{z}\|^2$ is strictly convex (because $\|\mathbf{X}\mathbf{z}\|^2 = \mathbf{z}^\top\mathbf{X}^\top\mathbf{X}\mathbf{z}$, where $\mathbf{X}^\top\mathbf{X} \succ 0$ since $\mathbf{X}^\top\mathbf{X}$ is assumed non-singular). By the definition of strict convexity, this means:

$$\|\theta\mathbf{X}\mathbf{u} + (1-\theta)\mathbf{X}\mathbf{v}\|^2 < \theta\|\mathbf{X}\mathbf{u}\|^2 + (1-\theta)\|\mathbf{X}\mathbf{v}\|^2$$

Therefore we can substitute the expression on the right into the previous equation to obtain:

$$\kappa(\theta\mathbf{u} + (1-\theta)\mathbf{v}) > \frac{2\theta\mathbf{y}^\top\mathbf{X}\mathbf{u} + 2(1-\theta)\mathbf{y}^\top\mathbf{X}\mathbf{v}}{\|\mathbf{y}\|^2 + \theta\|\mathbf{X}\mathbf{u}\|^2 + (1-\theta)\|\mathbf{X}\mathbf{v}\|^2} \tag{A-3}$$

We can then substitute the expression for $\mathbf{y}^\top\mathbf{X}\mathbf{v}$ from the right side of (A-2) into the above, leading to:

$$\begin{aligned}
&\frac{2\theta\mathbf{y}^\top\mathbf{X}\mathbf{u} + 2(1-\theta)\mathbf{y}^\top\mathbf{X}\mathbf{v}}{\|\mathbf{y}\|^2 + \theta\|\mathbf{X}\mathbf{u}\|^2 + (1-\theta)\|\mathbf{X}\mathbf{v}\|^2} \geq \\
&\qquad \frac{2\theta\mathbf{y}^\top\mathbf{X}\mathbf{u} + 2(1-\theta)\mathbf{y}^\top\mathbf{X}\mathbf{u}\left(\frac{\|\mathbf{y}\|^2+\|\mathbf{X}\mathbf{v}\|^2}{\|\mathbf{y}\|^2+\|\mathbf{X}\mathbf{u}\|^2}\right)}{\|\mathbf{y}\|^2 + \theta\|\mathbf{X}\mathbf{u}\|^2 + (1-\theta)\|\mathbf{X}\mathbf{v}\|^2} \\
&= \frac{2\mathbf{y}^\top\mathbf{X}\mathbf{u}\left(\theta + (1-\theta)\left(\frac{\|\mathbf{y}\|^2+\|\mathbf{X}\mathbf{v}\|^2}{\|\mathbf{y}\|^2+\|\mathbf{X}\mathbf{u}\|^2}\right)\right)}{\|\mathbf{y}\|^2 + \theta\|\mathbf{X}\mathbf{u}\|^2 + (1-\theta)\|\mathbf{X}\mathbf{v}\|^2} \\
&= \frac{2\mathbf{y}^\top\mathbf{X}\mathbf{u}\left(\frac{\theta(\|\mathbf{y}\|^2+\|\mathbf{X}\mathbf{u}\|^2)+(1-\theta)(\|\mathbf{y}\|^2+\|\mathbf{X}\mathbf{v}\|^2)}{\|\mathbf{y}\|^2+\|\mathbf{X}\mathbf{u}\|^2}\right)}{\|\mathbf{y}\|^2 + \theta\|\mathbf{X}\mathbf{u}\|^2 + (1-\theta)\|\mathbf{X}\mathbf{v}\|^2} \\
&= \frac{\frac{2\mathbf{y}^\top\mathbf{X}\mathbf{u}}{\|\mathbf{y}\|^2+\|\mathbf{X}\mathbf{u}\|^2}(\|\mathbf{y}\|^2 + \theta\|\mathbf{X}\mathbf{u}\|^2 + (1-\theta)\|\mathbf{X}\mathbf{v}\|^2)}{\|\mathbf{y}\|^2 + \theta\|\mathbf{X}\mathbf{u}\|^2 + (1-\theta)\|\mathbf{X}\mathbf{v}\|^2} \\
&= \kappa(\mathbf{u}) \quad .
\end{aligned}$$

Combining this result with (A-3) we have

$$\kappa(\theta\mathbf{u} + (1-\theta)\mathbf{v}) > \kappa(\mathbf{u})$$

or, stated another way

$$\kappa(\theta\mathbf{u} + (1-\theta)\mathbf{v}) > min[\kappa(\mathbf{u}), \kappa(\mathbf{v})]$$

which is what we were trying to prove. □

References

Avriel, Mordecai. *Nonlinear programming: analysis and methods*. Courier Corporation, 2003.