1 Introduction

We now turn our attention to not only estimating probabilities of default for each class, but also all transition probabilities among the classes as well, i.e. probabilities of transitioning from any given class to any other class, including remaining in the same class. This will lead to a transition matrix, with each row being associated with a given class and containing the the transition probabilities from the given class to all classes. We will assume the transition probabilities of any given class follow a multinomial distribution, and that these multinomial random variates are independent of each other. In estimating the transition matrix, we will also assume certain order restrictions among the transistion probabilities and hence enforce constraints on the parameter space.

2 Transition Matrix

In what follows, we will assume that there are k rating classes and denote these as $r_i, i = 1, \ldots, k$. The k^{th} class will be implicitly assumed to be the default class. Since the default state is an absorbing state which cannot be transitioned out of, the k^{th} row of the transition matrix need not be estimated, but will always consist of the vector $(0, 0, \ldots, 0, 1)^T$. Thus our transition matrix will take this form, with p_{ij} being the transition probability from class i to class j:

$$\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & p_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ p_{(k-1)1} & p_{(k-1)2} & \cdots & p_{(k-1)k} \\ 0 & 0 & \ldots & 1 \end{bmatrix}$$

Now, let n_i be the number of creditors in class i at a given time, and let y_{ij} be the number of these n_i that have transitioned to class j after a specified time period(e.g. one year). We can model these transitions from class i with a multinomial probability distribution as:

$$P_{r_i}(Y_{i1} = y_{i1}, \ldots, Y_{ik} = y_{ik}) = \frac{n_i!}{y_{i1}! \cdots y_{ik}!} p_{i1}^{y_{i1}} \cdots p_{ik}^{y_{ik}}$$

Assuming $P_{r_1}, \ldots, P_{r_{k-1}}$ are all independent of each other, and recalling the k^{th} class need not be estimated since it is the absorbing default state, we can write the unconstrained likelihood function as:

$$L(\mathbf{p}|\mathbf{n}, \mathbf{y}) = \prod_{i=1}^{k-1} \left(n_i \prod_{j=1}^{k} (y_{ij}!)^{-1} p_{ij}^{y_{ij}} \right)$$

where $\mathbf{n} = (n_1, \ldots, n_{k-1}), \mathbf{y} = (y_{11}, \ldots, y_{1k}, y_{21}, \ldots, y_{(k-1)k})$ and $\mathbf{p} = (p_{11}, \ldots, p_{1k}, p_{21}, \ldots, p_{(k-1)k})$

Then we can write the unconstrained log-likelihood, after ignoring constants, as:

$$l(\mathbf{p}) = \sum_{i=1}^{k-1} \sum_{j=1}^{k} y_{ij} \log(p_{ij})$$

It is well known that the maximum likelihood estimator (mle) of $\mathbf{p}$ is $\hat{\mathbf{p}} = (\hat{p}_{11}, \ldots, \hat{p}_{1k}, \hat{p}_{21}, \ldots, \hat{p}_{(k-1)k})$, where $\hat{p}_{ij} = \frac{y_{ij}}{n_i}$. Instead of using the mle for each parameter however, we will restrict the parameter space with constraints based on empirical observation as well as theoretical consideration. First, we note that for any given class a creditor is most likely to remain in the same class. Thus $p_{ii} \geq p_{ij}$ for $j = 1, \ldots, i-1, i+1, \ldots, k$. Next, for any given class, the probability of transitioning to another class decreases with the distance from the given class(with the usual absolute value distance metric). Combining this with the first set of inequalities above, this leads to a partial order on each row, consisting of two simple orders, in which we have: $p_{ii} \geq p_{i(i+1)} \geq \cdots \geq p_{ik}$ and $p_{ii} \geq p_{i(i-1)} \geq \cdots \geq p_{i1}$. Each row of the transition matrix then has an *umbrella* order with peak on the diagonal element: $p_{i1} \leq p_{i2} \leq \cdots \leq p_{ii} \geq p_{i(i+1)} \geq \cdots \geq p_{ik}$.

The row constraints deal with transition probabilities *from* a given class. We also introduce "symmetric" constraints that deal with transition probabilities *to* a given class. With reasoning similar to above, we note that the probability of remaining in the same state should be greater than the probability of transitioning into that state from any other state, hence $p_{jj} \geq p_{ij}$ for each i. Also, the probability of transitioning into a given state decreases with the distance from that state. Combining once again, we get an umbrella order on each row, with peak again at the diagonal element: $p_{1j} \leq p_{2j} \leq \cdots \leq p_{jj} \geq p_{(j+1)j} \geq p_{(k-1)j}$.

When estimating the transition matrix, we exclude the last row since it is already determined and extraneous to the problem, and focus on the $(k-1) \times k$ submatrix. For this submatrix, there are $(k-1)$ constraints on each row, and $(k-2)$ constraints on each column, so we have $(k-1)^2+k(k-2)$ constraints on $\mathbf{p}$, each in the form of a pairwise inequality. In addition, each of the $k-1$ rows is contrained to sum to 1. For the pairwise inequalities, we can represent any one of these by $\mathbf{a}^T\mathbf{p} \geq 0$, where $\mathbf{a}$ is a vector of 0s except for a 1 at the index of the greater element of the inequality and -1 at the index of the lesser element of the inequality. Thus all of the inequalities can be represented by $\mathbf{Ap} \geq \mathbf{0}$, where the rows of $\mathbf{A}$ are permutations of the vector $(1, -1, 0, \ldots, 0)$. It follows that $\mathbf{A}$ is of dimension $k(k-1)^2(k-2) \times k(k-1)$ and the parameter space is restricted to: $\{\mathbf{p}|\mathbf{Ap} \geq \mathbf{0}\}$. Of course we also have the individual multinomial constraints $\sum_{j=1}^{k} p_{ij} = 1$, for $i = 1, \ldots, k-1$.

In the case of a single multinomial distribution, finding the mle of $\mathbf{p} = (p_1, \ldots, p_k)$ subject to $\sum_i p_i = 1$ and $\mathbf{Ap} \geq \mathbf{0}$, where $\mathbf{A}$ is again defined by rows that are permutations of the vector $(1, -1, 0, \ldots, 0)$, can be reduced to a least squares problem. Let $(Y_1, \ldots, Y_k)$ follow a multnomial distribution with probability parameter $\mathbf{p} = (p_1, \ldots, p_k)$, and suppose we observe $Y_1 = y_1, \ldots, Y_k = y_k$. Let $z_i = y_i / \sum_{i=1}^{k} y_i$ and $w_i = y_i/z_i$. Silvapulle and Sen() (see also Barlowe() and Robertson et al.()) show that:

$$\tilde{\mathbf{p}} = \underset{\mathbf{Ap} \geq \mathbf{0}}{\arg\min} \sum_{i=1}^{k} (z_i - p_i)^2$$

is also the mle of $\mathbf{p}$ subject to $\mathbf{Ap} \geq \mathbf{0}$ and $\sum p_i = 1$. In the least squares problem above, the parameter space is only restricted by $\mathbf{Ap} \geq \mathbf{0}$, not $\sum_i p_i = 1$. This is because any solution to the least squares problem above must satisfy $\sum_i (z_i - \tilde{p}_i) = 0$. Since $\sum_i z_i = 1$, we must also have $\sum_i \tilde{p}_i = 1$. So this restraint is built in to the solution of the least squares. Now, noting that z_i is the unrestricted mle of p_i, it is tempting to try something similar in our case of multi-multinomials. If we let $y_{i\cdot} = \sum_j y_{ij}$, for $i = 1, \ldots, k$ and $z_{ij} = y_{ij}/y_{i\cdot}$, the unrestricted mle of p_{ij}, we might try to reduce our constrained nonlinear optimization problem to a least squares problem as in the case of a single multinomial. If we set:

$$\tilde{\mathbf{p}} = \underset{\mathbf{Ap} \geq \mathbf{0}}{\arg\min} \sum_{i=1}^{k-1} \sum_{j=1}^{k} (z_{ij} - p_{ij})^2$$

we again must have (details omitted, but can be adapted from Silvapulle and Sen()): $\sum_{i=1}^{k-1}\sum_{j=1}^{k}(z_{ij}-\tilde{p}_{ij})=0$. Unfortunately, this only requires $\sum_i\sum_j\tilde{p}_{ij}=k-1$, not the $k-1$ individual constraints $\sum_j p_{ij}=1$ we require. We cannot, therefore, reduce our nonlinear likelihood optimization to a least squares problem. Instead, the matrix $\mathbf{A}$ must be augmented to inlcude $k-1$ rows, $\mathbf{a_i}^T$, where $\mathbf{a_i}^T$ is the vector containing 1s in the corresponding k component indices that $p_{i1},\ldots,p_{ik}$ has in $\mathbf{p}$, and 0s elsewhere. Then this will give the equality constraints needed: $\mathbf{a_i}^T\mathbf{p}=1, i=1,\ldots,k-1$. We will still refer to this augmented matrix as $\mathbf{A}$ since it will now be the focus of our attention. Our problem, now, is to find $\tilde{\mathbf{p}}$ such that:

$$\tilde{\mathbf{p}}=\underset{\mathbf{Ap}\geq\mathbf{b}}{\arg\min}-\sum_{i=1}^{k-1}\sum_{j=1}^{k}y_{ij}\log(p_{ij}) \tag{1}$$

where $\mathbf{b}$ is now a vector of 0s except for the last $k-1$ coordinates, which equal 1. Note also we multiplied by -1 in order to represent this in the more standard form of a minimization problem.

3 Estimation of the Constrained Transition Matrix

We first note that since the transition matrix consists of rows of multinomial probabilities, each independent of the others, the solution to the unconstrained minimum of the negative log-likelihood corresponds to the ML estimates for each multinomial. Without the pairwise constraints then, for a given row i, we would have $\tilde{p}_{ij}=y_{ij}/\sum_{j=1}^{k}y_{ij}$. It follows that if we compute these ML estimates and they satisfy all the pairwise constraints, then this will be the solution to equation (1), the constrained mle. If there are a large number of observations, and the justifications for the constraints are valid, this should not be extremely unlikely. However, if we observe one or more constraint violations among these ML estimates, we must consider a different strategy. We therefore consider the Lagrangian framework for (1). To this end, we again define the log-likelihood, with the definining multinomial constraints, as:

$$l(\mathbf{p}) = \sum_{i=1}^{k-1}\sum_{j=1}^{k} y_{ij}\log(p_{ij}) \qquad 0 \le p_{ij} \le 1;\ \sum_{j=1}^{k} p_{ij} = 1,\ i = 1, \ldots, k-1$$

Then, using slightly different notation then (1), the problem is:

$$\begin{aligned} \min_{\mathbf{p}\in[0,1]^{k(k-1)}} \quad & -l(\mathbf{p}) \\ \text{subject to} \quad & \sum_{j=1}^{k} p_{ij} - 1 = 0 \qquad i = 1, \ldots, k-1 \\ & \mathbf{a_i}^T\mathbf{p} \ge 0 \qquad\qquad i = 1, \ldots, (k-1)^2 + k(k-2) \end{aligned} \tag{2}$$

where the $(k-1)^2 + k(k-2)$ vectors $\mathbf{a_i}^T$ are the pairwise constraint vectors from the matrix $\mathbf{A}$ defined above. We can then define the Lagrangian function $\mathcal{L}(\mathbf{p}, \lambda)$ as:

$$\mathcal{L}(\mathbf{p}, \lambda) = -l(\mathbf{p}) - \sum_{i=1}^{k-1} \lambda_i \left(\sum_{j=1}^{k} p_{ij} - 1 \right) - \sum_{m=k}^{(k-1)^2+k(k-2)+k-1} \lambda_m (\mathbf{a}_{(m-k+1)}^T \mathbf{p})$$

Since $-l(\mathbf{p})$ is convex and we are minimizing over a convex set (a unit hypercube with linear inequality contraints and affine equality constraints), any local minimizer is a global minimizer and is unique. The Karush-Kuhn-Tucker(KKT) conditions are thus necessary and sufficient for the global minimum. Let $\mathbf{p}^*$ and λ^* be such a minimum and associated multipliers, respectively. Then the KKT conditions are:

$$\nabla_{\mathbf{p}} \mathcal{L}(\mathbf{p}^*, \lambda^*) = 0 \tag{3a}$$

$$\sum_{j=1}^{k} p_{ij}^* - 1 = 0 \qquad i = 1, \ldots, k-1 \tag{3b}$$

$$\mathbf{a_i}^T\mathbf{p}^* \ge 0 \qquad i = 1, \ldots, (k-1)^2 + k(k-2) \tag{3c}$$

$$\lambda_i^* \ge 0 \qquad \text{for each } i \tag{3d}$$

$$\lambda_m(\mathbf{a}_{(m-k+1)}^T \mathbf{p}^*) = 0 \qquad m = k, \ldots, (k-1)^2 + k(k-2) + k - 1 \tag{3e}$$

If we find $(\mathbf{p}^*, \lambda^*)$ satisfying these conditions, then $\mathbf{p}^* = \tilde{\mathbf{p}}$, our constrained mle. Condition (3e) is referred to as the complentary slackness condition. Consider a typical pairwise inequality constraint, $p_{ij} \geq p_{kl}$ and let λ_m be the corresponding Lagrange multiplier. Then complentary slackness implies we must have $\lambda_m(p_{ij} - p_{kl}) = 0$; this further implies that if this constraint is *inactive*, i.e. p_{ij} is strictly greater than p_{kl}, then λ_m is identically 0. If a particular probability estimate, p_{ij} does not violate any of the pairwise inequality constraints, and hence all multipliers associated with those constraints vanish, then the gradient equation (3a) associated with that estimate will simply be $y_{ij}/p_{ij} - \lambda_i = 0$, λ_i being the multiplier associated with row i summing to 1. If there are no constraint violations at all, then we must have $\lambda_i(\sum_j p_{ij}) = \sum_j y_{ij} \Rightarrow \lambda_i = \sum_j y_{ij}$, which gives us $\tilde{p}_{ij} = y_{ij}/\sum_j y_{ij}$, the mle, as expected.

This, of course, is the most simple case. Let us know consider some other simple cases in which we have at least one pairwise inequality constraint violation. In what follows, row violations will refer to constraint violations on a given row, i.e. between p_{ij} and $p_{i(j+1)}$ for some i, j. Column violations will refer to a constraint violations between two elements on a given column, i.e. between p_{ij} and $p_{(i+1)j}$ for some i, j.

3.1 Row Constraint Violations Only

After generating the unconstrained ML estimates, suppose we have only row constraints. For a given row i, note first that each side of the diagonal element p_{ii} consists of a simple order. For any violation of a simple order, we have previously seen that the Pool Adjacent Violators Algorithm (PAVA) may be used to arrive at an optimal constrained solution. There is then only one other case to consider, that of a double violation of the diagonal element, $p_{i(i-1)} > p_{ii} < p_{i(i+1)}$. We could use the Lagrangian equations to establish a solution, but it is easy to see with a simple example. Without loss of generality, assume $p_{21} = .2, p_{22} = .1$, and $p_{23} = .4$. The question is which violation to reconcile first. If we choose the "lesser" violation first, and use PAVA, we get $p_{21} = p_{22} = .15$, so we will still have the second, "greater", violation since $p_{22} = .15 < .4 = p_{23}$. Pooling again, we get $p_{21} = p_{22} = p_{23} = .333...$. On the other hand, if we use PAVA on the "larger" violation first, then we get $p_{22} = p_{23} = .25$. Now there are no violations since this change in p_{22} had the effect of annuling the "lesser" violation. There are cases when choosing the "larger" violator will still end in the three estimates

being pooled, and this can be made precise with the Lagrange multipliers. However, it should be clear that the latter method of choosing the "larger" violator will always end in the correct solution, whereas choosing the lesser violator will always give the same solution, which is sometimes erroneous.

To summarize, if there are only row violations from the ML estimates, we can use a modified PAVA to find the optimal constrained transition matrix. The modified PAVA uses regular PAVA for any simple order violation and adds the method of choosing the larger violator in the case of a double violation with the diagonal element.

3.2 One Column Violation Only

We now assume that among the ML estimates, the only constraint violation is a column constraint. Without loss of generality, assume $p_{11} < p_{21}$. If we pool and equalize these, clearly the rows will no longer sum to 1. Thus to equalize these, we will need to redistribute the change in "weight" of these elements among the remaining row elements to ensure the row sums to 1. We will bring in the Lagrangian machinery to determine precisely how to make the constraint violation feasible and redistribute the weights of the rows. Let λ_m be the multiplier associated with the p_{11}, p_{21} constraint. The relevant Lagrangian gradient equations are:

$$\begin{aligned}
\frac{y_{11}}{p_{11}} - \lambda_1 - \lambda_m &= 0 \\
\frac{y_{12}}{p_{12}} - \lambda_1 &= 0 \\
&\vdots \\
\frac{y_{1k}}{p_{1k}} - \lambda_1 &= 0 \\
\frac{y_{21}}{p_{21}} - \lambda_2 - \lambda_m &= 0 \\
\frac{y_{22}}{p_{22}} - \lambda_2 &= 0 \\
&\vdots \\
\frac{y_{2k}}{p_{2k}} - \lambda_1 &= 0
\end{aligned} \tag{4}$$

Setting $p_{21} = p_{11}$ and adding the equations involving λ_m, we get:

$$\frac{y_{11} + y_{21}}{p_{11}} - \lambda_1 - \lambda_2 = 0 \tag{5}$$

Next, summing up the remaining equations involving λ_1 and using the constraint $\sum_{i=2}^{k} p_{1i} = 1 - p_{11}$, we get:

$$\lambda_1 = \frac{\sum_{i=2}^{k} y_{1i}}{1 - p_{11}} \tag{6}$$

And similarly,

$$\lambda_2 = \frac{\sum_{i=2}^{k} y_{2i}}{1 - p_{11}} \tag{7}$$

Now substituting (6) and (7) into (5), we have:

$$\frac{y_{11} + y_{21}}{p_{11}} = \frac{\sum_{i=2}^{k} y_{1i} + y_{2i}}{1 - p_{11}}$$

Finally, solving, we have:

$$p_{11} = p_{21} = \frac{y_{11} + y_{21}}{\sum_{i=1}^{k} (y_{1i} + y_{2i})}$$

and:

$$\lambda_1 = \frac{\left[\sum_{i=1}^{k} (y_{1i} + y_{2i})\right] \left[\sum_{i=2}^{k} y_{1i}\right]}{\sum_{i=2}^{k} (y_{1i} + y_{2i})}$$

$$\lambda_2 = \frac{\left[\sum_{i=1}^{k} (y_{1i} + y_{2i})\right] \left[\sum_{i=2}^{k} y_{2i}\right]}{\sum_{i=2}^{k} (y_{1i} + y_{2i})}$$

Note that $p_{1i} = y_{1i}/\lambda_1$ and $p_{2i} = y_{2i}/\lambda_2$ for $i = 1, \ldots, k$. Thus the non-constraint-violating probability estimates are all reweighted by the same amount.

So if we only have one column constraint violation, we can pool the violating elements and reweight the remaining row elements by the reciprocal of their corresponding row Lagrange multiplier.

3.3 Row Constraint Violations with One Column Constraint Violation

From above, we know that a column constraint violation equalizes the violating constraint and reweights the remaining elements of each row by a specific row weight, e.g. $1/\lambda_i$ for row i. Since the elements of a row are all reweighted by the same weight, any order on the row will still hold after reweighting. It follows we can use the two strategies above in combination if we have a column constraint violation along with row constraint violations on the rows involved. One additional consideration is necessary: if we have one probability estimate involved in a column violation *and* a row violation, we choose the "largest" violation to pool first, for identical reasons to the modified PAVA case mentioned for row constraint violators.

3.4 Two Column Constraint Violations between Three Consecutive Rows

Unfortunately, the search for a simple algorithmical solution to the estimation of the transition matrix under the constraints given above becomes much more difficult when two column constraints are violated on three consecutives rows. If the violations are not all in the same column (in which case a relatively simple solution is available) then the Lagrangian equations become a system of nonlinear equations that are very difficult to solve analytically. In such cases, numerical procedures such as an augmented Lagrangian numerical procedure would be required.

References

[1] Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). Statistical Inference Under Order Restrictions. John Wiley and Sons, New York

[2] Robertson, T., Wright, F.T., Dykstra, R.L., 1988. Order Restricted Statistical Inference. Wiley.

[3] Silvapulle, M., Sen, P., 2005. Constrained Statistical Inference. Wiley-Interscience.