Wypasek Data Science, Inc.

Thoughts on the concurrent fitting/combining of default and prepayment and matrix model fitting

Documentation

Version 0.0.2

Author(s): Christian Wypasek

Abstract

Christian Wypasek's Notes, 2022: These are some really old thoughts on fitting competing risk models.

The concepts in this document were originally conceived and implemented in late 1999 and 2000. I had pursued possibly academically publishing it, however, upon further research, Sir David Cox himself had proposed the core concepts in the 1980s. Between 2004 and 2005, I developed a simple techinque to trick SAS's PHReg into fitting competing risks, and have used and further developed that basic methodology ever since. Further developments included the use of progressive conal ordering models for robustness and a baseline fitting method based on martigale residuals. Details will be in WDS-ModelSpec docs.

Interestingly enough, after a few years without a SAS license, it came to my attention that some of the concepts were published by Fine and Gray in 1999 and some proportional hazards modeling software now have "Fine and Gray" extensions. Based on Cox's precedence, I would argue that is the term should only be *Cox Proportional Hazards for Competing Risks*.

1 Current Techniques

Obviously, in order to understand cash flow, both prepayment and default need to be well understood. Whereas default is (hopefully) a small proportion of any overall pool with a small impact on prepayment, the prepayment rate constantly changes the dynamic of default propensity of the surviving pool. The marginal effect of a late cycle default on default timing estimates becomes inflated. In the prime world, this may potentially be due to non-informative censoring witnessed by evidence that prepayment speed tends to be negatively correlated to default risk.

The techniques we have used or are considering have all tried to adjust for this phenomenon and may be described as one of the following:

1. **Mark then life (MTL)**
   The intuition is that the ultimate default propensity exists prior to life. In this scenario, to find $C^p(t)$, the cumulative prepayment probability, one must first find $C^d(\infty)$, the ultimate default probability, and then find $C^p(t)$ via the product of $1 - C^d(\infty)$ and the conditional probability of prepayment by time t given the loan will never go into default.

   The challenge remains the overestimation of the default rate and the solution technique is to treat prepayments as non-defaults throughout the study.

2. **Competing lives with fastest receiving the mark (a.k.a. the racing game)**

   The intuition is that often the marginal default timing is inflated and similarly the prepayment timing may be slightly inflated. Compensate these two effects by "racing" the two estimated distributions, i.e., Let X be the time to default and let Y be the time to prepayment and let $S^d(t)$ and $S^p(t)$ be their respective survival functions. Then our estimate of $C^d(t)$ becomes:

   $$C^d(t) = \mathrm{P}\left(X \leq t, X < Y\right) = \int_0^t -\mathrm{P}\left[X \leq y | Y = y\right] dS^p(y) + \mathrm{P}\left(X \leq t, Y > t\right),$$

   which under the assumption of independence of X and Y (at least given the common covariate picture) becomes:

   $$C^d(t) = \int_0^t -\mathrm{P}\left(X \leq y\right) dS^p(y) + (1 - S^d(t))S^p(t).$$

   The challenge remains the overestimation of the default rate and the solution technique is to allow this transformation of S^d and S^p to damp out the biasing of their estimates. Even though there is a damping out of the biasing of $C^d(t)$ especially at $t = \infty$ by correcting for prepayment, both prepay and default curves may be faster than intended. Furthermore, the effectiveness may depend on the relative speeds and $C^d(\infty)$.

   CJW–The more I know and the more I have seen through examples, the more I realize that there are still too many effects confounding the overall timing with this technique. At GE, this seemed intuitive and seemed to work, but through work with MTL and LTM I think we now know why there are problems and how to correct them.

3. **Life then mark (LTM)**

   The intuition is the loan has a "productive life" the end of which carries a mark of prepayment or default. Consider the graph above and let $S(t)$ denote the total survival function and let $h(t), h^d(t), h^p(t)$ represent the hazard rates for the total survival, default and prepayment respectively. It follows that $S(t)$ satisfies the differential equation:

   $$\frac{dS(t)}{dt} = -h(t)S(t) = -(h^d(t) + h^p(t))S(t).$$

   Furthermore,

   $$\frac{dC^d(t)}{dt} = h^d(t)S(t) \text{ and } \frac{dC^p(t)}{dt} = h^p(t)S(t).$$

   Note that the estimation of $S(t)$ and $h(t)$ can be performed knowing that the censoring is truely non-informative. Under some absolute continuity assumptions, let $\nu(t) = \frac{h^d(t)}{h^p(t)}$, the instantaneous odds of a default at time t. The result becomes:

$$\frac{dC^d(t)}{dt} = \frac{\nu(t)}{1+\nu(t)} h(t) S(t).$$

The derivative ν can be estimated via a separate but interesting (logistic) model for the odds of a default given that the non-productive time occurs at time t. Here, t is treated as just another variable, but its inclusion redirects the exit patterns, i.e., early exits may have a higher default rate followed by a prepay exodus until burnout followed by steady state leaving rates.

For conditional models as in the case where a loan has not gone to default or prepayment after time t_1, the differential equation can just be solved forward with $S(t_1) = 1$. This will also incorporate only the exit patterns which exist after time t_1.

The above technique for fitting LTM is the "fast" way and easy to fit in SAS. However, the estimation of $S(t)$ may be compounded by the fact that it is a mixture of both default and prepayment timing distributions. Any variables used to fit $S(t)$ are equally applied to both parts of the mixture. The "long(er)" process for fitting both the derivative and the total survival proceeds below.

2 First, Some Light Background Discussion

CJW-Please forgive the following fault: I tend to move rather freely between discrete time and continuous time. Motivations tend to be in continuous time, then applied discrete time with a subjectively small Δt. This should not be much of a problem in a non-Brownian motion stochastic calculus. (Ok, I have found places like integrating out the baselines that need to be done in a discretized manner and not just rectangular on the theoretical, otherwise the discussion still flows.)

For a nicely behaved lifetime, X, with distribution function, F, and density, f, the hazard rate is defined as:

$$h(t) = \frac{f(t)}{1 - F(t)}.$$

F can then be written:

$$F(t) = 1 - e^{-\int_0^t h(u)du}.$$

Conditional survival distributions are taken as:

$$\mathrm{P}\left(X \leq t | X > c\right) = 1 - e^{-\int_c^t h(u)du}.$$

And for small increments, Δt,

$$\mathrm{P}\left(X \leq t + \Delta t | X > t\right) \approx h(t)\Delta t.$$

In proportional hazards models, for a given covariate picture, Z, the forms for h in continuous and discrete times are taken as:

$$h(t) = g(\beta Z) h_0(t) \qquad \text{and} \qquad \frac{h(t)\Delta t}{1 - h(t)\Delta t} = \frac{h_0(t)\Delta t}{1 - h_0(t)\Delta t} g(\beta Z).$$

Here, h_0 is a baseline hazards rate that is not parameterized but can be estimated from the data, and g is a function, usually $g(x) = e^x$ to ensure non-negativity. Essentially, given a baseline rate, a proportional hazards is a small increment proportional odds model (notice the similarities to $\frac{\nu(t)}{1+\nu(t)} h(t)$ above).

Fitting a proportional hazard rate model derives from Poisson processes. Suppose a collection of Poisson processes, $N_t^1, \ldots, N_t^n$, with intensities, $\lambda_1, \ldots, \lambda_n$, is aggregated, i.e., $N_t = \sum_{i=1}^n N_t^i$. Furthermore, suppose an event for N occurs at time t, then the probability that the event occurs due to Poisson process i is given by:

$$\frac{\lambda_i}{\sum_{j=1}^n \lambda_j.}$$

Given a collection of covariate pictures, $Z_1, \ldots, Z_N$, and event times $T_1, \ldots, T_N$, (assume times are ordered, $T_i < T_j$ when $i < j$), the likelihood for subject i is:

$$\frac{h_i(t)}{\sum_{j=i}^N h_j(t)} = \frac{e^{\beta Z_i} h_0(t)}{\sum_{j=i}^N e^{\beta Z_j} h_0(t)} = \frac{e^{\beta Z_i}}{\sum_{j=i}^N e^{\beta Z_j}},$$

i.e., the baseline cancels out and the likelihood is the proportional factor over the sum of factors for all subjects able to fail at time T_i. In general, this is only considered a partial likelihood for the dynamics surrounding $h_0(t)$ are left un-specified, hence it is also a semi-parametric model. This interpretation also facilitates right censoring in which case the denominator above includes all subjects able to fail at time T_i, but only subjects that actually fail contribute to the overall likelihood of the model/dataset.

The nature of the likelihood function is inherently linked to the Kaplan-Meier estimator, a variant of which is used to estimate $S_0(t) = e^{-\int_0^t h(u)du}$. Once $S_0(t)$ is estimated from the data, $S_i(t) = [S_0(t)]^{e^{\beta Z_i}}$.

3 Working Marked Point Processes into the Mix

3.1 Fitting the long technique

Instead of playing a racing game with composite distributions as above in technique 2 which introduced scores of censoring/numerical/technique issues, let us look to the loan level. Instead of treating subject i as a single counting process stream as described in the previous section, treat each subject as a composite stream itself, i.e., $N_t^i = N_t^{i,d} + N_t^{i,p}$ (even though we generally care only about the first event epoch). Note that this is exactly a 3-state case of a semi-Markov renewal model (I need to check Cinlar for the exact name I'm looking for) which is like a continuous time Markov chain except sojourn times are general distributions. In other words, this type of procedure could be used for fitting timed-multinomials, possibly conditional on a given state with the multinomial outcome as the transistion into another state.

As before, assume that event times are ordered, $T_1, \ldots, T_N$, with a 1-1 mapping to subjects and let $M_1, \ldots, M_N$, be the respective marks which may be d, p or e for default, prepayment or end of study respectively. For our purposes end of study times are the only purposefull but truely non-informative censoring times for we will assume there are no other stopping times. As before censored observations do not have unique contributions to likelihood. Let $h_i^d(t) = e^{\beta^d Z_i} h_0^d(t), h_i^p(t) = e^{\beta^p Z_i} h_0^p(t)$ be the default and prepayment hazards for subject i. For a non-censored observation, i, the contribution to likelihood is:

$$\frac{1_{\{M_i=d\}} e^{\beta^d Z_i} h_0^d(T_i) + 1_{\{M_i=p\}} e^{\beta^p Z_i} h_0^p(T_i)}{\sum_{j=i}^N e^{\beta^d Z_j} h_0^d(T_i) + e^{\beta^p Z_j} h_0^p(T_i)}.$$

Here, $1_{\{A\}}$ is the indicator of a set, i.e., $1_{\{A\}} = 1$ if A occurs and 0 otherwise.

Of course, it would be nice if we could *cancel* out the baseline hazards, but they are not shared in common any longer. However, dividing numerator and denominator by $h_0^d(T_i)$ yields a contribution to likelihood of :

$$\frac{1_{\{M_i=d\}} e^{\beta^d Z_i} + 1_{\{M_i=p\}} e^{\beta^p Z_i} \frac{h_0^p(T_i)}{h_0^d(T_i)}}{\sum_{j=i}^N e^{\beta^d Z_j} + e^{\beta^p Z_j} \frac{h_0^p(T_i)}{h_0^d(T_i)}}.$$

Here, $\frac{h_0^p(t)}{h_0^d(t)}$ is the instanteous odds ratio of prepayments to defaults at time t as before in LTM. Now, take a standard logistic form of $\frac{h_0^p(t)}{h_0^d(t)} = e^{\phi(t)}$. Note that this ratio (actually a Radon-Nikodym derivative between the two measures and therefore induces some absolute continuity assumptions between the default and prepayment support) can be found from the definition of ν in the LTM discussion above. $\nu(t)$ can be estimated from a separate odds model of the form $\nu(t) = e^{\hat{\beta} Z_i + \phi(t)}$ fit on loans with non-censored observations. The term $\hat{\beta} Z_i$ becomes a controlling factor to find the average effect of time alone. In a general sense, there may be terms of Z_i which are correlated to time and could be included above, but this would become more computationally intensive and we will not consider that immmediately. To see this complexity, notice that we are dividing by $h_0^d(T_i)$ at the time of the numerators event, interactions with time would require the denominators covariates to be interacted with the numerators time for each T_i.

The process for fitting the marked point process becomes as follows:

- Fit the derivative effect, $\phi(t)$ via a logistic model, $\nu(t) = e^{\hat{\beta} Z_i + \phi(t)}$.

- Maximize the following likelihood function:

$$\prod_{M_i \neq e} \frac{1_{\{M_i=d\}} e^{\beta^d Z_i} + 1_{\{M_i=p\}} e^{\beta^p Z_i} e^{\phi(T_i)}}{\sum_{j=i}^{N} e^{\beta^d Z_j} + \left(\sum_{j=i}^{N} e^{\beta^p Z_j}\right) e^{\phi(T_i)}}.$$

Of course, it looks like $\hat{\beta} = \beta^p - \beta^d$, however, these sub-models can be entirely specified as to what is appropriate for prepayment or what is appropriate for default. Furthermore, these coefficients are fit in a non-informative censoring environment.

3.2 Estimating the baseline

Most references on baselines either give a formula or some glancing idea at how it was derived, but to understand its extension, it helps to see the derivation. Even though it is commonly said that the baseline hazard rate is estimated empirically from the data, it is actually a maximum likelihood estimate. Assume that the event times $T_1, \ldots, T_N$ are ordered as above, but let us temporarily assume that there are no censored observations. The baseline will be estimated in a multiplicative manner, and let α_n be the factor:

$$\alpha_n = \frac{S_0(T_n)}{S_0(T_{n-1})}.$$

It follows that $S(T_n)$ has the form:

$$S(T_n) = \left[\prod_{1 \leq k \leq n} \alpha_k\right]^{e^{\beta Z_n}}.$$

The probability mass function then has the form:

$$pmf_n = (1 - \alpha_n^{e^{\beta Z_n}}) \prod_{1 \leq k < n} \alpha_k^{e^{\beta Z_n}}.$$

Given β, the total likelihood of the dataset with parameters α_k is given by the product of corresponding terms of the above form, the log-likelihood is:

$$\log \prod_{1 \leq n \leq N} pmf_n = \sum_{1 \leq n \leq N} \log(1 - \alpha_n^{e^{\beta Z_n}}) + \sum_{1 \leq n \leq N} \sum_{1 \leq k < n} e^{\beta Z_n} \log(\alpha_k).$$

To maximize this log-likelihood, let us look to the point where the derivatives of the above quantity with respect to each α_n equate to zero. These derivatives are given by:

$$\frac{\delta LL}{\delta \alpha_n} = \frac{1}{1 - \alpha_n^{e^{\beta Z_n}}} \left[-e^{\beta Z_n} \alpha_n^{e^{\beta Z_n} - 1}\right] + \frac{1}{\alpha_n} \sum_{k > n} e^{\beta Z_k}.$$

Equating to 0 for each $1 \leq n \leq N$, and solving for the term, $\alpha_n^{e^{\beta Z_n}}$, one obtains:

$$\alpha_n^{e^{\beta Z_n}} = \frac{e^{-\beta Z_n} \sum_{k>n} e^{\beta Z_k}}{1 + e^{-\beta Z_n} \sum_{k>n} e^{\beta Z_k}}.$$

A quick gyration yields:

$$\alpha_n = \left[1 - \frac{e^{\beta Z_n}}{\sum_{k \geq n} e^{\beta Z_k}}\right]^{e^{-\beta Z_n}}.$$

This solution holds when each subject with an observed event as a unique event time. The equation is adapted in the case of ties and the technique is to solve a system of equations. If this is not easily observable, consider the following: Suppose there is a multiplicity of events at time n, say l_n. If every Z_i with $t_i = t_n$ were identical, adding a factor of l_n to the derivatives above, setting equal to 0 and solving would be elementary. However, the summation of $\alpha_n^{e^{\beta Z_i}}$ terms complicates this.

Extending this concept to censoring is simple, only non-censored observations contribute a term to the likelihood. Extending this to marked point processes follows exactly the same concept, if a loan prepays, its default stream is censored and only the prepayment term contributes to likelihood. There will also be switching from continuous to discrete time and back again to obtain the proceeding algorithm. **If one would want to save oneself considerable grief, with these insights one could skip to the final answer of this section and understand the result.** Otherwise,....

In continuous time, the total survival function for subject n is given by:

$$S^n(t) = 1 - e^{-\int_0^t \left(e^{\beta^d Z_n} + e^{\beta^p Z_n} e^{\phi(u)}\right) h_0(u) du}.$$

For defaults, the cumulative distribution satisfies the following differential equation:

$$\frac{dC^d(t)}{dt} = e^{\beta^d Z_n} h_0(t) S^n(t),$$

and for prepayments:

$$\frac{dC^p(t)}{dt} = e^{\beta^p Z_n} e^{\phi(t)} h_0(t) S^n(t),$$

Given the mark of subject n, M_n, and taking the time at t_n, let $\Delta t_n = t_n - t_{n-1}$ and let us rewrite these two densities in a way amenable to the probability mass function above:

$$pmf_n \approx \left[1_{\{M_n=d\}} e^{\beta^d Z_n} + 1_{\{M_n=p\}} e^{\beta^p Z_n} e^{\phi(t_n)}\right] h_0(t_n) e^{-\int_{t_{n-1}}^{t_n} \left(e^{\beta^d Z_n} + e^{\beta^p Z_n} e^{\phi(u)}\right) h_0(u) du} S^n(t_{n-1}) \Delta t_n.$$

To make this equation a little less unwieldy, let us temporarily use the notation, $a_n = e^{\beta^d Z_n}$, $b_n = e^{\beta^p Z_n}$, and $c(t) = e^{\phi(t)}$, rewriting the equation as,

$$pmf_n \approx \left[1_{\{M_n=d\}} a_n + 1_{\{M_n=p\}} b_n c(t_n)\right] h_0(t_n) e^{-\int_{t_{n-1}}^{t_n} (a_n + b_n c(t_n)) h_0(u) du} S^n(t_{n-1}) \Delta t_n.$$

Taking $c(t) = t_n$ and $h_0(t) = h_0(t_n)$ over $(t_{n-1}, t_n]$, and recalling that $\lambda \Delta t \approx 1 - e^{-\lambda \Delta t}$ for an exponential, simplifies the expression to

$$\begin{aligned}
pmf_n &\approx \left[1_{\{M_n=d\}} a_n + 1_{\{M_n=p\}} b_n c(t_n)\right] h_0(t_n) e^{-(a_n + b_n c(t_n)) h_0(t_n) \Delta t_n} S^n(t_{n-1}) \Delta t_n. \\
&\approx \Big[1_{\{M_n=d\}} a_n h_0(t_n) \Delta t_n e^{-a_n h_0(t_n) \Delta t_n} e^{-b_n c(t_n) h_0(t_n) \Delta t_n} \\
&\qquad + 1_{\{M_n=p\}} b_n c(t_n) h_0(t_n) \Delta t_n e^{-b_n c(t_n) h_0(t_n) \Delta t_n} e^{-a_n h_0(t_n) \Delta t_n}\Big] S^n(t_{n-1}). \\
&\approx \Big[1_{\{M_n=d\}} (1 - e^{-a_n h_0(t_n) \Delta t_n}) e^{-b_n c(t_n) h_0(t_n) \Delta t_n} \\
&\qquad + 1_{\{M_n=p\}} (1 - e^{-b_n c(t_n) h_0(t_n) \Delta t_n}) e^{-a_n h_0(t_n) \Delta t_n}\Big] S^n(t_{n-1}).
\end{aligned}$$

Taking $\alpha_n = e^{-h_0(t_n) \Delta t_n}$, yields:

$$\begin{aligned}
pmf_n &\approx \left[1_{\{M_n=d\}} (1 - \alpha_n^{a_n}) \alpha_n^{b_n} + 1_{\{M_n=p\}} (1 - \alpha_n^{b_n c(t_n)}) \alpha_n^{a_n}\right] S^n(t_{n-1}). \\
&\approx \left[1_{\{M_n=d\}} (1 - \alpha_n^{a_n}) \alpha_n^{b_n} + 1_{\{M_n=p\}} (1 - \alpha_n^{b_n c(t_n)}) \alpha_n^{a_n}\right] \prod_{1 \le k < n} \alpha_n^{a_n + b_n c(t_n)}.
\end{aligned}$$

Following the technique above for a vanilla baseline calculation, taking the derivative of the log-likelihood yields:

$$\begin{aligned}
\frac{\delta LL}{\delta \alpha_n} &= 1_{\{M_n=d\}} \left[\frac{1}{1 - \alpha_n^{a_n}} (-a_n \alpha_n^{a_n - 1}) + \frac{b_n}{\alpha_n}\right] \\
&\quad + 1_{\{M_n=p\}} \left[\frac{1}{1 - \alpha_n^{b_n c(t_n)}} (-b_n c(t_n) \alpha^{b_n c(t_n) - 1}) + \frac{a_n}{\alpha_n}\right] \\
&\quad + \frac{1}{\alpha_n} \sum_{k > n} (a_k + b_k c(t_n)).
\end{aligned}$$

It may take a moment of thought to convince oneself, but within the last summation for the k^{th} term, $c(t)$ must be evaluated at t_n. Equating each partial derivative to zero and using a similar gyration as above, one obtains:

$$\begin{aligned}\alpha_n &= \left[1 - \frac{1_{\{M_n=d\}}a_n + 1_{\{M_n=p\}}b_n c(t_n)}{\sum_{k\geq n} a_k + b_k c(t_n)}\right]^{\frac{1}{1_{\{M_n=d\}}a_n + 1_{\{M_n=p\}}b_n c(t_n)}} \\ &= \left[1 - \frac{1_{\{M_n=d\}}e^{\beta^d Z_n} + 1_{\{M_n=p\}}e^{\beta^p Z_n}(t_n)}{\sum_{k\geq n} e^{\beta^d Z_k} + e^{\beta^p Z_k}e^{\phi(t_n)}}\right]^{\frac{1}{1_{\{M_n=d\}}e^{\beta^d Z_n} + 1_{\{M_n=p\}}e^{\beta^p Z_n}e^{\phi(t_n)}}}.\end{aligned}$$

In the case of tied event times, the system of equations generated by setting derivatives to 0 is solved.

4 Further extensions and equations for Newton-Raphson fitting

Once the format for two outcomes is established, extension to a greater number of outcomes is direct. For example, if the mark set is $m_0, m_1, \ldots, m_l$, (with m_0 as the censored mark), the key to this extension is a family of Radon-Nikodym derivatives, $\phi_2(t), \ldots, \phi_l(t)$, between the multiple hazard functions and one pre-specified reference outcome. These could of course be fit using a common reference generalized logit model. The quanitities above of the form $1_{\{M_n=d\}}e^{\beta^d Z_n} + 1_{\{M_n=p\}}e^{\beta^p Z_n}e^{\phi(t_n)}$ would be replaced by

$$1_{\{M_n=m_1\}}e^{\beta^1 Z_n} + \sum_{k=2}^{l} 1_{\{M_n=m_k\}}e^{\beta^k Z_n}e^{\phi_k(t_n)}.$$

If we use a linear form for $\phi_k(t) = \phi_k t$, (now treating ϕ_k and t as vector valued), the ϕ^k could also be fit simulataneously to the β^k. This would alleviate some of the censored data issues around the preconditioning model. Let us consider the following form for the likelihood function:

$$L = \prod_{M_i \neq e} \frac{\sum_{k=1}^{l} 1_{\{M_n=m_k\}}e^{\beta^k Z_i}e^{\phi_k T_{i,i}}\eta(i,i)}{\sum_{j=i}^{N}\left(\sum_{k=1}^{l} e^{\beta^k Z_j}e^{\phi_k T_{j,i}}\right)\eta(i,j)}.$$

In case you missed it, there is a ϕ_1. Let us generalize T_j to $T_{j,i}$, where $T_{j,i}$ represents a vector of time varying covariates for observation j taken at the time of observation i. By appropriate selection of variables, one can use time varying data such as incentive or home price index even for the reference class. There is an open question of identifiability. For time in the usual sense, this issue can be resolved by coding the problem in a fully flexible manner where one can totally specify which variables get used for each model, i.e., time in the usual sense will not be specified for the reference class. Recall that one of the reasons that time is a special case is that we are using the Radon-Nikodym derivatives to eliminate *as much as is common to all members of the cohort at that event epoch.* Therefore any variable that has the same impact for all observations must be treated as a time-like variable. Observation unique or at least some-what unique variables can be specified as time-varying.

Let us also consider the strata functional $\eta(i,j)$. It will generally be assumed that $\eta(i,i) = 1$ but it is included for completeness. In the usual sense of *strata* in a Cox model, η will be either 0 or 1 depending on whether or not i and j are within the same strata. Let us generalize this to the sense of a potentially *soft* strata case where η is a neighbor defining metric. It can also be a catch all term for a weighting scheme. Clearly bayesian notions will surround the parameter definitions within η and the area will need some more thought.

Admittedly, it was first noticed during early modeling experiments, but there is a natural interaction between the concepts of the previous two paragraphs. A hard strata structure is used whereever one works under the assumption of a fixed proportionality function for each covariate, regardless of the baseline. However, the strata structure and functional, $\eta(i,j)$, are used to guarantee that only common baselines are canceled out. It is therefore only natural to discover that if different baselines are reasonable for different cohorts, then the Radon-Nikodym derivatives, ϕ^k, are also likely to change by strata. In our generalized covariate, T, we may want to consider some elements treated as *time-like* variables in the usual sense which probably need to be interacted with strata, or maybe not, in order to fit a fixed proportionality function across cohorts. Once again, a fully flexible piece of software for fitting will cure the problem.

For model fitting, we will employ Newton's method, for a quick reminder: Let our vector of parameters be $\vec{\beta}$. Since the likelihood L is the product of probabilities (usually small) and these often have a multiplicative or exponential form themselves, we will maximize the log-likelihood, LL, for an additive form. The candidate for a maximum must satisfy $\nabla LL(\vec{\beta}) = 0$. If the Fisher information matrix, $-\mathrm{E}\left[\nabla^2 LL(\vec{\beta})\right]$, is nicely behaved, i.e., nothing like quasi-complete separation in logistic regression, the following algorithm should converge to a maximum: Given an initial estimate, $\vec{\beta}_0$, iterate until desired convergence through $\vec{\beta}_{n+1} = \vec{\beta}_n + \Delta\vec{\beta}_n$ where

$$\nabla^2 LL(\vec{\beta}_n)\Delta\vec{\beta}_n = -\nabla LL\vec{\beta}_n.$$

For testing variable significance, we will use the inverse of the Fisher information matrix as a variance estimate of the parameter estimates,

$$cov(\vec{\beta}) = (-\nabla^2 LL(\vec{\beta}))^{-1}.$$

Thus, for a Newton-Raphson technique the following quantities are required:

$$\frac{\delta LL}{\delta\beta_a^{k_1}} = \sum_{M_i\neq m_0}\left[1_{\{M_n=m_{k_1}\}}Z_{i,a}\eta(i,i) - \frac{\sum_{j=i}^N Z_{j,a}e^{\beta^{k_1}Z_j}e^{\phi_{k_1}T_{j,i}}\eta(i,j)}{\sum_{j=i}^N\sum_{k=1}^l e^{\beta^k Z_j}e^{\phi_k T_{j,i}}\eta(i,j)}\right]$$

$$\frac{\delta LL}{\delta\phi_a^{k_1}} = \sum_{M_i\neq m_0}\left[1_{\{M_n=m_{k_1}\}}T_{i,i,a}\eta(i,i) - \frac{\sum_{j=i}^N T_{j,i,a}e^{\beta^{k_1}Z_j}e^{\phi_{k_1}T_{j,i}}\eta(i,j)}{\sum_{j=i}^N\sum_{k=1}^l e^{\beta^k Z_j}e^{\phi_k T_{j,i}}\eta(i,j)}\right]$$

$$\frac{\delta^2 LL}{\delta\beta_a^{k_1}\delta\beta_b^{k_1}} = \sum_{M_i\neq m_0}\left[-\frac{\sum_{j=i}^N Z_{j,a}Z_{j,b}e^{\beta^{k_1}Z_j}e^{\phi_{k_1}T_{j,i}}\eta(i,j)}{\sum_{j=i}^N\sum_{k=1}^l e^{\beta^k Z_j}e^{\phi_k T_{j,i}}\eta(i,j)} + \frac{\sum_{j=i}^N Z_{j,a}e^{\beta^{k_1}Z_j}e^{\phi_{k_1}T_{j,i}}\sum_{j=i}^N Z_{j,b}e^{\beta^{k_1}Z_j}e^{\phi_{k_1}T_{j,i}}\eta(i,j)}{\left(\sum_{j=i}^N\sum_{k=1}^l e^{\beta^k Z_j}e^{\phi_k T_{j,i}}\eta(i,j)\right)^2}\right]$$

$$\frac{\delta^2 LL}{\delta\beta_a^{k_1}\delta\phi_b^{k_1}} = \sum_{M_i\neq m_0}\left[-\frac{\sum_{j=i}^N Z_{j,a}T_{j,i,b}e^{\beta^{k_1}Z_j}e^{\phi_{k_1}T_{j,i}}\eta(i,j)}{\sum_{j=i}^N\sum_{k=1}^l e^{\beta^k Z_j}e^{\phi_k T_{j,i}}\eta(i,j)} + \frac{\sum_{j=i}^N Z_{j,a}e^{\beta^{k_1}Z_j}e^{\phi_{k_1}T_{j,i}}\eta(i,j)\sum_{j=i}^N T_{j,i,b}e^{\beta^{k_1}Z_j}e^{\phi_{k_1}T_{j,i}}\eta(i,j)}{\left(\sum_{j=i}^N\sum_{k=1}^l e^{\beta^k Z_j}e^{\phi_k T_{j,i}}\eta(i,j)\right)^2}\right]$$

$$\frac{\delta^2 LL}{\delta\phi_a^{k_1}\delta\phi_b^{k_1}} = \sum_{M_i\neq m_0}\left[-\frac{\sum_{j=i}^N T_{j,i,a}T_{j,i,b}e^{\beta^{k_1}Z_j}e^{\phi_{k_1}T_{j,i}}\eta(i,j)}{\sum_{j=i}^N\sum_{k=1}^l e^{\beta^k Z_j}e^{\phi_k T_{j,i}}\eta(i,j)} + \frac{\sum_{j=i}^N T_{j,i,a}e^{\beta^{k_1}Z_j}e^{\phi_{k_1}T_{j,i}}\eta(i,j)\sum_{j=i}^N T_{j,i,b}e^{\beta^{k_1}Z_j}e^{\phi_{k_1}T_{j,i}}\eta(i,j)}{\left(\sum_{j=i}^N\sum_{k=1}^l e^{\beta^k Z_j}e^{\phi_k T_{j,i}}\eta(i,j)\right)^2}\right]$$

$$\frac{\delta^2 LL}{\delta\beta_a^{k_1}\delta\beta_b^{k_2}} = \sum_{M_i\neq m_0}\left[\frac{\sum_{j=i}^N Z_{j,a}e^{\beta^{k_1}Z_j}e^{\phi_{k_1}T_{j,i}}\eta(i,j)\sum_{j=i}^N Z_{j,b}e^{\beta^{k_2}Z_j}e^{\phi_{k_2}T_{j,i}}\eta(i,j)}{\left(\sum_{j=i}^N\sum_{k=1}^l e^{\beta^k Z_j}e^{\phi_k T_{j,i}}\eta(i,j)\right)^2}\right]$$

$$\frac{\delta^2 LL}{\delta\beta_a^{k_1}\delta\phi_b^{k_2}} = \sum_{M_i\neq m_0}\left[\frac{\sum_{j=i}^N Z_{j,a}e^{\beta^{k_1}Z_j}e^{\phi_{k_1}T_{j,i}}\eta(i,j)\sum_{j=i}^N T_{j,i,b}e^{\beta^{k_2}Z_j}e^{\phi_{k_2}T_{j,i}}\eta(i,j)}{\left(\sum_{j=i}^N\sum_{k=1}^l e^{\beta^k Z_j}e^{\phi_k T_{j,i}}\eta(i,j)\right)^2}\right]$$

$$\frac{\delta^2 LL}{\delta\phi_a^{k_1}\delta\phi_b^{k_2}} = \sum_{M_i\neq m_0}\left[\frac{\sum_{j=i}^N T_{j,i,a}e^{\beta^{k_1}Z_j}e^{\phi_{k_1}T_{j,i}}\eta(i,j)\sum_{j=i}^N T_{j,i,b}e^{\beta^{k_2}Z_j}e^{\phi_{k_2}T_{j,i}}\eta(i,j)}{\left(\sum_{j=i}^N\sum_{k=1}^l e^{\beta^k Z_j}e^{\phi_k T_{j,i}}\eta(i,j)\right)^2}\right]$$

Let's talk a minute about fitting the baseline, (this discussion should be above but we'll work that around later). In the case of no ties and regarding the LL for the baseline and talking the product below only over non-censored observations, we have

$$LL = \log\prod_{1\le n\le N} pmf_n = \sum_{1\le n\le N}\left[\sum_{k=1}^l 1_{\{M_n=m_k\}}\log(1-\alpha_n^{e^{\beta^k Z_n}e^{\phi^k T_{n,n}}}) + \sum_{1\le j<n}\left(\sum_{k=1}^l e^{\beta^k Z_n}e^{\phi^k T_{n,j}}\right)\log(\alpha_j)\right].$$

Recall that the parameters, α_k, are the factors of the baseline survival. For the two outcome case, we had the following derivative (in a slightly simpler form than presented previously):

$$\frac{\delta LL}{\delta\alpha_n} = \frac{1}{\alpha_n}\left(-1_{\{M_n=d\}}\frac{a_n}{1-\alpha_n^{a_n}} - 1_{\{M_n=p\}}\frac{b_n c(t_n)}{1-\alpha_n^{b_n c(t_n)}} + \sum_{k\ge n}(a_k + b_k c(t_n))\right).$$

In the case of ties and applying the multinomial form above, we obtain,

$$\frac{\delta LL}{\delta \alpha_n} = \frac{1}{\alpha_n}\left(-\sum_{\{t_g=t_n \cap M_g \neq m_0\}}\sum_{k=1}^{l} 1_{\{M_g=m_k\}}\frac{e^{\beta^k Z_g e^{\phi_k T_{g,n}}}}{1-\alpha_n^{e^{\beta^k Z_g e^{\phi_k T_{g,n}}}}} + \sum_{t_g \geq t_n}\sum_{k=0}^{l} e^{\beta^k Z_g e^{\phi_k T_{g,n}}}\right).$$

With some abuse of notation, the above assumes only one α_n for each unique time t_n. The second derivative proceeds directly,

$$\frac{\delta^2 LL}{\delta \alpha_n^2} = \frac{1}{\alpha_n}\left(-\frac{\delta LL}{\delta \alpha_n} - \sum_{\{t_g=t_n \cap M_g \neq m_0\}}\sum_{k=1}^{l} 1_{\{M_g=m_k\}}\frac{\left(e^{\beta^k Z_g e^{\phi_k T_{g,n}}}\right)^2 \alpha_n^{e^{\beta^k Z_g e^{\phi_k T_{g,n}}}-1}}{\left(1-\alpha_n^{e^{\beta^k Z_g e^{\phi_k T_{g,n}}}}\right)^2}\right).$$

It follows that the second derivative is negative whenever the first derivative is 0 and at least a local maximum criterion is satisfied.

The other more pressing reason to consider the the first and second derivatives is a parameterization of the baseline. Now, it is clear that anything where the first derivative above is non-zero is sub-optimal, however, a parameterization cleans up the baseline and reduces overfitting noise. Since for each α_n there is a unique t_n, suppose that we have the form $\alpha_n = \theta(t_n)$ and that $\theta(t)$ has a parameter θ_p. Note that it is required that $0 \leq \theta(t) \leq 1$. We can then use the following equations for Newton-Raphson fitting:

$$\begin{aligned}
\frac{\delta LL}{\delta \theta_p} &= \sum_n \frac{\delta LL}{\delta \alpha_n}\frac{\delta \alpha_n}{\delta \theta_p} \\
\frac{\delta^2 LL}{\delta \theta_p^2} &= \sum_n \left[\frac{\delta LL^2}{\delta \alpha_n^2}\left(\frac{\delta \alpha_n}{\delta \theta_p}\right)^2 + \frac{\delta LL}{\delta \alpha_n}\frac{\delta^2 \alpha_n}{\delta \theta_p^2}\right] \\
\frac{\delta^2 LL}{\delta \theta_p \delta \theta_q} &= \sum_n \left[\frac{\delta LL^2}{\delta \alpha_n^2}\frac{\delta \alpha_n}{\delta \theta_p}\frac{\delta \alpha_n}{\delta \theta_q} + \frac{\delta LL}{\delta \alpha_n}\frac{\delta^2 \alpha_n}{\delta \theta_p \delta \theta_q}\right]
\end{aligned}$$

5 References

1. *Applied Survival Analysis, Regression Modeling of Time to Event Data*, David W. Hosmer, Jr. and Stanly Lemeshow, Wiley Series in Probability and Statistics, 1999.
2. Cinlar

6 Discrete-time parameterized geometric matrix models

Here we are using the term discrete-time parameterized geometric matrix models to describe an entire family models that can be constructed via state-wise conditioning methods. Sometimes directly, sometimes indirectly, this family actually encompasses:

- Ordinary (binary) logistic regression
- Multinomial logistic regression
- DMMs, or more appropriately named discrete time Markov chains (DTMCs, usually stationary)
- Net-flow models (under-determined or partially hidden DTMCs)
- DDMMs, or non-stationary DTMCs
- Discrete-time proportional hazards with competing risks via a geometric, full likelihood construction

For this discussion of this type of model, let us first start with the form of the dataset. This will motivate the discussion of the form of the model and the likelihood construction. The likelihood is motivated by logistic regression and we will review that case briefly. The form of the likelihood is very natural and the maximum likelihood estimates with Fisher information by-products can be found via a Newton-Raephson technique. The required first and second derivative expressions are included below.

6.1 Dataset form

The original motivation for this discussion comes from estimating net-flow roll matrices from aggregate pool-level information, such as Intex data. Net-flow models are naturally considered since tracking each dollar in a pool is usually not possible. However, since Net-flow models are special cases of DTMCs, it makes fortuitous sense to consider the most general case possible and limit ourselves only to what the data can support.

At each time point in a net-flow model, one has a distribution vector describing the current state of the system. The model is a roll matrix or linear tranformation of the current distribution vector to create an estimate of the expected distribution at the next time step. The error associated the with the transformation is garnered with the observed data at the next time step. With notation described below, the basic model form is:

$$\mathrm{E}\left[\widehat{\pi}(obs_i, t_{obs_i}+1)|obs_i, \beta\right] = \widehat{\pi}(obs_i, t_{obs_i})P(\beta, Z_{obs_i}).$$

Although the distribution information at the next time step may be redundantly stored on another observation, we will assume that the dataset has the basic form of:

$$\begin{pmatrix} obs_1 \\ obs_2 \\ \vdots \\ obs_N \end{pmatrix} = \begin{pmatrix} Z_{obs_1} & \widehat{\pi}(obs_1, t_{obs_1}) & \widehat{\pi}(obs_1, t_{obs_1}+1) \\ Z_{obs_2} & \widehat{\pi}(obs_2, t_{obs_2}) & \widehat{\pi}(obs_2, t_{obs_2}+1) \\ \vdots & & \\ Z_{obs_N} & \widehat{\pi}(obs_N, t_{obs_N}) & \widehat{\pi}(obs_N, t_{obs_N}+1) \end{pmatrix}$$

Here, for each observation, $obs_1, \ldots, obs_N$, there is an associated time point, t_{obs_1}, and covariate picture, Z_{obs_1}. The notation may seem slightly cumbersome at the moment, but only to emphasize that t_{obs_2} may or may not equal $t_{obs_1}+1$. With the understanding of the above form, the $obs_{\cdot}$ notation will usually be dropped.

With this form of the dataset, each time point creates a distinct observation that does not depend on any other time point. This facilitates combining different sources of data to which a common model may be estimated. We will not restrict the form of the covariate picture, Z, which may be time-dependent (including any time-like variables) and complex or simply an constant 1 for an intercept term (yes, if an intercept is desired it must be in the dataset and specified as an ordinary variable).

The distribution vector may or may not be normalized. If the distribution vector is pre-normalized, it is assumed that the total weight associated with each observiation is 1. We usually refer to non-normalized cases as events/trials that represent repeated measures experiments. The repeated measures may result from a pooled collection of loans and/or the dollars associated with a loan.

6.2 Logistic motivations of model form

Classic binary logistic regression has the following properties: each observation represents the one trial of a Bernoulli experiment upon a unique subject. The binary response is usually called a *good* or a *bad*. It is assumed that the dataset represents N independent trials for which the probability of one response (say good) is bernoulli with probablity perfectly determined by:

$$\mathrm{P}\left(\text{good}\right) = p = \frac{e^{\beta Z}}{1+e^{\beta Z}}$$

The dataset usually looks like:

$$\begin{pmatrix} obs_1 \\ obs_2 \\ \vdots \\ obs_N \end{pmatrix} = \begin{pmatrix} Z_{obs_1} & 0 \text{ or } 1 \\ Z_{obs_2} & 0 \text{ or } 1 \\ \vdots & \\ Z_{obs_N} & 0 \text{ or } 1 \end{pmatrix}$$

where the response is 1 for a good and 0 for a bad.

However, this can be translated into the form of the previous section by realizing that this is a two state model, calling the states 1 and 2. The trial response of a good represents starting in state 1 and moving to state 2 in one time step and a response of a bad comes from staying in state 1 during the trial. Here, we set $\widehat{\pi}(t) = [1\ 0]$ for $t = 0$ and $\widehat{\pi}(1) = [1\ 0]$ for a bad and $\widehat{\pi}(1) = [0\ 1]$ for a good. The transistion matrix model is very simple:

$$\begin{pmatrix} \frac{1}{1+e^{\beta Z}} & \frac{e^{\beta Z}}{1+e^{\beta Z}} \\ 0 & 1 \end{pmatrix}$$

The likelihood function for the logistic case can then be written as follows:

$$L(\beta) = \prod_{obs} (\widehat{p}_1(1))^{\widehat{\pi}_1(1)} (\widehat{p}_2(1))^{\widehat{\pi}_2(1)}$$

where $\widehat{p}(1) = \mathrm{E}\left[\widehat{\pi}(1)|Z\right]$, $\widehat{p}_1(1) = \frac{1}{1+e^{\beta Z}}$ and $\widehat{p}_2(1) = 1 - \widehat{p}_1(1)$. One benefit of this particular likelihood expression is that observations with identical covariate pictures can be grouped. For example, suppose each Z belongs to only one of m classes, $Z^1, \ldots, Z^m$. Let N^i denote the number of observations of class i and let $\widehat{\pi}^i$ denote the sum of $\widehat{\pi}$ where $Z = Z^i$. By rearranging the products above, it follows that:

$$\begin{aligned} L(\beta) &= \prod_{obs} (\widehat{p}_1(1))^{\widehat{\pi}_1(1)} (\widehat{p}_2(1))^{\widehat{\pi}_2(1)} \\ L(\beta) &= \prod_{i=1}^{m} \prod_{Z=Z^i} (\widehat{p}_1(1))^{\widehat{\pi}_1(1)} (\widehat{p}_2(1))^{\widehat{\pi}_2(1)} \\ L(\beta) &= \prod_{i=1}^{m} \prod_{Z=Z^i} (\widehat{p}^i_1(1))^{\widehat{\pi}_1(1)} (\widehat{p}^i_2(1))^{\widehat{\pi}_2(1)} \\ L(\beta) &= \prod_{i=1}^{m} (\widehat{p}^i_1(1))^{\widehat{\pi}^i_1(1)} (\widehat{p}^i_2(1))^{\widehat{\pi}^i_2(1)} \end{aligned}$$

This grouping of observations with identical covariate pictures is why the likelihood form for events/trials scenarios is exactly the same as if each element of an aggregate (loan in a pool or dollar in a loan/loan pool) was treated as a separate subject. There are dispersion issues to be taken into account, but we will set this temporarily aside for this discussion.

The multinomial form is a natural generalization. In this form there are n-states, 1 through n. The subject starts in state 1 and may transistion during 1 time step. The transistion matrix form for $n = 3$ is:

$$\begin{pmatrix} \frac{1}{1+e^{\beta_{12} Z}+e^{\beta_{13} Z}} & \frac{e^{\beta_{12} Z}}{1+e^{\beta_{12} Z}+e^{\beta_{13} Z}} & \frac{e^{\beta_{13} Z}}{1+e^{\beta_{12} Z}+e^{\beta_{13} Z}} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

The likelihood function for the multinomial case follows in a similar fashion:

$$L(\beta) = \prod_{obs} \prod_{j=1}^{n} (\widehat{p}_j(1))^{\widehat{\pi}_j(1)}.$$

It should be noted that the above structure also represents a discrete-time proportional hazards model with competing risks. The estimation method described here is a geometric, full likelihood construction, as opposed to a partial-likelihood construction as in Life-Then-Mark. Geometric construction implies that a loan is incorporated into the dataset, one observation for each surviving time point and one for the terminal event time. All observations for each loan are used in estimating the single-step, roll-into-event probability. This is in contrast to the partial-likelihood survival technique where the time-dependent roll-into-event rate uses each loan once against a comparable surviving set. Ultimately, via a monthly hazard rate, other competing risk survival models can be incorporated into the matrix model.

6.3 The matrix model likelihood

In a short, sweet summary, the discrete-time parameterized geometric matrix models are special cases of the multinomial logistic model just described. For each observation, the current distribution ($\widehat{\pi}(t)$) is a pre-disposing set of

values that combined with the transistion matrix, $P(\beta, Z)$, yields a new set of probabilities, $\widehat{p}(t+1)$ for a multinomial experiment. The likelihood function is:

$$L(\beta) = \prod_{obs} \prod_{j=1}^{n} (\widehat{p}_j(t+1))^{\widehat{\pi}_j(t+1)},$$

where

$$\widehat{p}(t+1) = \mathrm{E}\left[\widehat{\pi}(t+1)|\beta, Z\right] / \Sigma\widehat{\pi}(t) = \widehat{\pi}(t)P(\beta, Z)/\Sigma\widehat{\pi}(t).$$

Note: the $\Sigma\widehat{\pi}(t)$ is required for the number of events in events/trials scenarios and it is assumed that there is mass conservation.

With that done, let us move on to estimating β.

6.4 Estimating β

Notation is going to become unusually ugly, so let us adopt the following conventions and notations:

- For indexing, we will always use the following letters in the same manner:

Rows	Columns	Covariate Index
i	j	k
a	b	c
u	v	w

- The actual number of covariates is left unspecified and treated as a vector $Z = (Z_k)$ for each observation.
- It will probably be common that the transistion matrix is sparse, but in keeping with a total general discussion, we will assume it is densely populated, $n \times n$ matrix: $P(\beta, Z) = (P_{ij}(\beta, Z))_{1 \le i,j \le n}$.
- Transistion matrices must have row sums of 1. To parameterize this correctly, we treat each row as a sub-model with the diagonal element as the *reference state*. In other words, with n columns for each row, there are $n-1$ degrees of freedom, each with its own sub-sub-model, and the 1 degree of freedom lost is associated with the diagonal element. The coefficient vector is a triply-indexed set, $\beta = (\beta_{ijk})_{1 \le i,j \le n, i \ne j}$, but we will use $\beta_{ij}Z$ to represent $\Sigma\beta_{ijk}Z_k$. and

$$P_{ii}(\beta, Z) = \frac{1}{1 + \sum_{1 \le v \le n, v \ne i} e^{\beta_{iv}Z}}$$
$$p_{ij}(\beta, Z) = \frac{e^{\beta_{ij}Z}}{1 + \sum_{1 \le v \le n, v \ne i} e^{\beta_{iv}Z}}$$

As before, the expected response for each observation is given by:

$$\widehat{p}(t+1) = \mathrm{E}\left[\widehat{\pi}(t+1)|\beta, Z\right] / \Sigma\widehat{\pi}(t) = \widehat{\pi}(t)P(\beta, Z)/\Sigma\widehat{\pi}(t)$$

and the likehood function is

$$L(\beta) = \prod_{obs} \prod_{j=1}^{n} (\widehat{p}_j(t+1))^{\widehat{\pi}_j(t+1)}.$$

The log-likelihood is used in practice,

$$LL(\beta) = \sum_{obs} \sum_{j=1}^{n} \widehat{\pi}_j(t+1) \log(\widehat{p}_j(t+1)).$$

For model fitting, we will employ Newton's method, for a quick reminder: Let our parameters, β be stacked into a vector, $\vec{\beta}$. The candidate for a maximum must satisfy $\nabla LL(\vec{\beta}) = 0$. If the Fisher information matrix, $-\mathrm{E}\left[\nabla^2 LL(\vec{\beta})\right]$, is nicely behaved, i.e., nothing like quasi-complete separation in logistic regression, the following

algorithm should converge to a maximum: Given an initial estimate, $\vec{\beta}_0$, iterate until desired convergence through $\vec{\beta}_{n+1} = \vec{\beta}_n + \Delta\vec{\beta}_n$ where

$$\nabla^2 LL(\vec{\beta}_n)\Delta\vec{\beta}_n = -\nabla LL\vec{\beta}_n.$$

For testing variable significance, we will use the inverse of the Fisher information matrix as a variance estimate of the parameter estimates,

$$cov(\vec{\beta}) = (-\nabla^2 LL(\vec{\beta}))^{-1}.$$

We therefore need the following quantities: $\frac{\delta LL(\beta)}{\delta\beta_{abc}}$ and $\frac{\delta^2 LL(\beta)}{\delta\beta_{abc}\delta\beta_{uvw}}$. These quantities can be related back to $\widehat{\pi}$, $\widehat{p}$, and $P(\beta, Z)$ through,

$$\begin{aligned}
\frac{\delta LL(\beta)}{\delta\beta_{abc}} &= \sum_{obs}\sum_{j=1}^{n} \frac{\widehat{\pi}_j(t+1)}{\widehat{p}_j(t+1)} \frac{\delta\widehat{p}_j(t+1)}{\delta\beta_{abc}} \\
\frac{\delta LL(\beta)}{\delta\beta_{abc}} &= \sum_{obs}\sum_{j=1}^{n} \frac{\widehat{\pi}_j(t+1)}{\widehat{p}_j(t+1)} \sum_{i=1}^{n} \widehat{\pi}_i(t) \frac{\delta P_{ij}(\beta, Z)}{\delta\beta_{abc}} \\
\frac{\delta^2 LL(\beta)}{\delta\beta_{abc}\delta\beta_{uvw}} &= \sum_{obs}\sum_{j=1}^{n} \frac{\widehat{\pi}_j(t+1)}{\widehat{p}_j(t+1)} * \\
&\qquad \left[-\frac{1}{\widehat{p}_j(t+1)} \left(\sum_{i=1}^{n} \widehat{\pi}_i(t) \frac{\delta P_{ij}(\beta, Z)}{\delta\beta_{abc}} \right) \left(\sum_{i=1}^{n} \widehat{\pi}_i(t) \frac{\delta P_{ij}(\beta, Z)}{\delta\beta_{uvw}} \right) + \sum_{i=1}^{n} \widehat{\pi}_i(t) \frac{\delta^2 P_{ij}(\beta, Z)}{\delta\beta_{abc}\delta\beta_{uvw}} \right]
\end{aligned}$$

The remaining quantities are $\frac{\delta P_{ij}(\beta,Z)}{\delta\beta_{abc}}$ and $\frac{\delta^2 P_{ij}(\beta,Z)}{\delta\beta_{abc}\delta\beta_{uvw}}$. After a small amount of grief, they can be show to be:

$$\begin{aligned}
\frac{\delta P_{ij}}{\delta\beta_{abc}} &= 1_{\{i=a\}} Z_c \left[-P_{ij}P_{ib} + 1_{\{j=b\}} P_{ij} \right] \\
\frac{\delta^2 P_{ij}}{\delta\beta_{abc}\delta\beta_{uvw}} &= 1_{\{i=a\}} 1_{\{i=u\}} Z_c Z_w \left[2P_{ij}P_{iv}P_{ib} - 1_{\{j=v\}} P_{iv}P_{ib} - 1_{\{j=b\}} P_{iv}P_{ib} - 1_{\{b=v\}} P_{ij}P_{ib} + 1_{\{j=b=v\}} P_{ij} \right]
\end{aligned}$$

(Here the dependence on (β, Z) is implicit.)