A Derivation of the 3-entity-type model

The 3-entity-type model consists of two relationship matrices, a $m \times n$ matrix X and a $n \times r$ matrix Y. In the spirit of generalized linear models we define the low rank representations of the relationship matrices to be $X \approx f_1(UV^T)$ and $Y \approx f_2(VZ^T)$ where $f_1 : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ and $f_2 : \mathbb{R}^{n\times r} \to \mathbb{R}^{n\times r}$ are the prediction links, and $U \in \mathbb{R}^{m\times k}$, $V \in \mathbb{R}^{n\times k}$, and $Z \in \mathbb{R}^{r\times k}$ are the parameters of the model for $k \in \mathbb{Z}$. We further define parameter transformations $G : \mathbb{R}^{m\times k} \to \mathbb{R}^{m\times k}$, $H : \mathbb{R}^{n\times k} \to \mathbb{R}^{n\times k}$, $I : \mathbb{R}^{r\times k} \to \mathbb{R}^{r\times k}$, to model prior knowledge about our parameters, *e.g.*, regularizers. We additionally required the convex conjugate,

$$G^*(U) = \sup_{U\in \mathrm{dom}(A)} \left[U \circ A - G(A)\right],$$

where $U \circ A = tr(U^T A) = \sum_{ij} U_{ij}A_{ij}$ is the matrix dot product. The overall loss function for our model is

$$L(U, V, Z|W, \tilde{W}) = \alpha L_1(U, V|W) + (1 - \alpha)L_2(V, Z|\tilde{W}) \tag{1}$$

where we introduce fixed weight matrices for the observations, $W \in \mathbb{R}^{m\times n}$ and $\tilde{W} \in \mathbb{R}^{n\times r}$. The individual objectives on the reconstruction of X and Y are, respectively,

$$L_1(U, V|W) = W \odot \left(F_1(UV^T) - X \circ UV^T\right) + G^*(U) + H^*(V), \tag{2}$$

$$L_2(V, Z\tilde{W}) = \tilde{W} \odot \left(F_2(VZ^T) - Y \circ VZ^T\right) + H^*(V) + I^*(Z). \tag{3}$$

The objective $L(U, V, Z|W, \tilde{W})$ is convex in any one of its arguments, but is in general non-convex in all its arguments. As such, we use an alternating minimization scheme that optimizes one factor U, V, or Z at a time. This appendix describes the derivation of both gradient and Newton update rules for U, V, and Z. For completeness, Section A.1 reviews useful definitions from matrix calculus. The gradient of the objective, with respect to each argument is derived in Section A.2. Finally, by assuming the loss is decomposable, we derive the Newton update in Section A.3, whose additional cost over its gradient analogue is essentially a factor of k times more expensive.

A.1 Matrix Calculus

For the sake of completeness we define matrix derivatives, which generalizes both scalar and vector derivatives. Using this definition of matrix derivatives, we also generalize the scalar chain and product rules to matrices. The discussion herein is based on Magnus et. al. [1, 2].

Let M be an $n \times q$ matrix of variables, where $m_{\cdot j}$ denotes the j-th column of M. The vec-operator $\mathrm{vec}\, M$ yields an $nq \times 1$ matrix that stacks the columns of M:

$$\mathrm{vec}\, M = \begin{pmatrix} m_{\cdot 1} \\ m_{\cdot 2} \\ \vdots \\ m_{\cdot q} \end{pmatrix}.$$

While there are several common (and incompatible) definitions of matrix derivatives, the derivative of a $n \times 1$ vector f with respect to a $m \times 1$ vector x is almost universally defined as

$$Df(x) \equiv \frac{\partial f}{\partial x^T} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_m} \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_m} \end{pmatrix}.$$

The matrix derivative of an $m \times p$ matrix function Q of an $n \times q$ matrix of variables M contains $mnpq$ partial derivatives, and the matrix derivative arranges these partial derivatives into a matrix. We define the matrix

derivative by coercing matrices into vectors, and using the above definition of the vector derivative,

$$DQ(M) \equiv \frac{\partial \operatorname{vec} Q(M)}{\partial (\operatorname{vec} M)^T} = \begin{pmatrix} \frac{\partial [Q(M)]_{11}}{\partial m_{11}} & \cdots & \frac{\partial [Q(M)]_{11}}{\partial m_{nq}} \\ \vdots & & \vdots \\ \frac{\partial [Q(M)]_{mp}}{\partial m_{11}} & \cdots & \frac{\partial [Q(M)]_{mp}}{\partial m_{nq}} \end{pmatrix},$$

which is an $mp \times nq$ matrix of partial derivatives. This definition encompasses vector and scalar derivatives as special cases. The advantages of this formulation include (i) unambiguous definitions for the product and chain rules, and (ii) we can easily convert $DQ(M)$ to the more common definition of the matrix derivative,

$$\frac{\partial Q(M)}{\partial M} = \begin{pmatrix} \frac{\partial Q(M)}{\partial m_{11}} & \cdots & \frac{\partial Q(M)}{\partial m_{1q}} \\ \vdots & & \vdots \\ \frac{\partial Q(M)}{\partial m_{n1}} & \cdots & \frac{\partial Q(M)}{\partial m_{nq}} \end{pmatrix},$$

via the first identification theorem [2, ch. 9]. We additionally require the matrix chain [2, pg. 121] and matrix product [1] rules:

Definition 1 (Matrix Chain Rule)**.** *Given functions $\varphi : \mathbb{R}^{n\times q} \to \mathbb{R}^{m\times p}$, $\varphi_1 : \mathbb{R}^{\ell\times r} \to \mathbb{R}^{m\times p}$, and $\varphi_2 : \mathbb{R}^{n\times q} \to \mathbb{R}^{\ell\times r}$ the derivative of $\varphi(M) = \varphi_1(Y)$, where $Y = \varphi_2(M)$, is*

$$D\varphi(Z) = D\varphi_1(Y) \times D\varphi_2(Z).$$

where $A \times B$ is the matrix product.

Definition 2 (Matrix Product Rule)**.** *Given an $m\times p$ matrix $\varphi_1(M)$ and a $p\times r$ matrix $\varphi_2(M)$ the derivative of $\varphi_1(M)\varphi_2(M)$ with respect to M, $D(\varphi_1\varphi_2)(M)$ is*

$$D(\varphi_1\varphi_2)(Z) = (\varphi_2(M)^T \otimes I_m) \cdot D\varphi_1(M) + (I_r \otimes \varphi_1(M)) \cdot D\varphi_2(M)$$

where $A \otimes B$ is the Kronecker product.

A.2 Computing the Gradient

To compute the derivative of Equation 1 with respect to U, V, and Z we require the following three lemmas:

Lemma 1. *For any differentiable function $F : \mathbb{R}^{m\times n} \to \mathbb{R}$*

$$\frac{\partial F_1(UV^T)}{\partial U} = f_1(UV^T)V, \quad \frac{\partial F_1(UV^T)}{\partial V} = f_1(UV^T)^T U$$
$$\frac{\partial F_2(VZ^T)}{\partial V} = f_2(VZ^T)Z, \quad \frac{\partial F_2(VZ^T)}{\partial Z} = f_2(VZ^T)^T V$$

Proof. We only derive the result for $\varphi(U) = F_1(UV^T)$. The proof is similar for the other three cases. $\varphi(U)$ can be expressed as the composition of functions: $\varphi(U) = F_1(Y)$, $Y = \varphi_2(U)$, $\varphi_2(U) = UV^T$. Using the matrix chain rule $D\varphi(U) = DF_1(Y) \cdot D\varphi_2(U)$, we note that

$$DF_1(Y) = \frac{\partial \operatorname{vec} F_1(UV^T)}{(\operatorname{vec} UV^T)^T} = f_1(UV^T). \tag{4}$$

Using the matrix product rule

$$\begin{aligned} D\varphi_2(U) &= (V \otimes I_m)\frac{\partial \operatorname{vec} U}{\partial (\operatorname{vec} U)^T} + (I_n \otimes U)\frac{\partial \operatorname{vec} V^T}{\partial (\operatorname{vec} U)^T} \\ &= (V \otimes I_m). \end{aligned} \tag{5}$$

Combining equations 4 and 5 using the matrix chain rule yields $D\varphi(U) = (\operatorname{vec} f_1(UV^T)V)^T$. The result follows immediately from the first identification theorem. □

Lemma 2. *Given that the entries of U and V are distinct,*

$$\frac{\partial(X \circ UV^T)}{\partial U} = XV, \quad \frac{\partial(X \circ UV^T)}{\partial V} = X^T U$$
$$\frac{\partial(Y \circ VZ^T)}{\partial V} = YZ, \quad \frac{\partial(Y \circ VZ^T)}{\partial Z} = Y^T V.$$

Proof. We derive the result for $\partial(X \circ UV^T)/\partial V$, the other three derivations are similar. To avoid a long digression into matrix differentials, we prove the result by element-wise differentiation. Noting that

$$X \circ UV^T = \sum_i \sum_j x_{ji} \sum_k u_{jk} v_{ik},$$

we compute the derivative with respect to v_{pq}

$$\begin{aligned}\frac{\partial(X \circ UV^T)}{\partial v_{pq}} &= \sum_i \sum_j x_{ji} \frac{\partial}{\partial v_{pq}} \sum_k u_{jk} v_{ik} \\ &= \sum_j x_{jp} u_{jq} = (X^T U)_{pq}\end{aligned}$$

Since the result holds for all $p \in \{1, \ldots, n\}$ and $q \in \{1, \ldots, k\}$ it follows that $\partial(X \circ UV^T)/\partial V = X^T U$. □

Lemma 3. *For the ℓ_2 regularizers where $a, b, c > 0$ controls the strength of regularizers (larger values $\implies$ weaker regularization):*

$$G(U) = \frac{a||U||^2_{Fro}}{2}, \quad H(V) = \frac{b||V||^2_{Fro}}{2}, \quad I(Z) = \frac{c||Z||^2_{Fro}}{2},$$

the derivatives for the convex conjugates are

$$\frac{\partial G^*(U)}{\partial U} = \frac{U}{a} = A, \quad \frac{\partial H^*(V)}{\partial V} = \frac{V}{b} = B, \quad \frac{\partial I^*(Z)}{\partial Z} = \frac{Z}{c} = C.$$

Proof. The result is easily proven by finding the convex conjugate and differentiating it. □

Combining Lemmas 1,2, and 3, and denoting the Hadamard (element-wise) product of matrices $A \odot B$, we have that

$$\frac{\partial L(U,V,Z)}{\partial U} = \alpha \left(W \odot \left(f_1(UV^T) - X\right)\right) V + A, \tag{6}$$

$$\frac{\partial L(U,V,Z)}{\partial V} = \alpha \left(W \odot \left(f_1(UV^T) - X\right)\right)^T U + (1-\alpha) \left(\tilde{W} \odot \left(f_2(VZ^T) - Y\right)\right) Z + B, \tag{7}$$

$$\frac{\partial L(U,V,Z)}{\partial Z} = (1-\alpha) \left(\tilde{W} \odot \left(f_2(VZ^T) - Y\right)\right)^T V + C. \tag{8}$$

Setting the gradient equal to zero yields update equations either for A, B, C or for U, V, Z. An advantage of using a gradient update is that we can relax the requirement that the links are differentiable, replacing gradients with subgradients in the prequel.

A.3 Computing the Newton Update

One may be satisfied with a gradient step. However, the assumption of a decomposable loss means that most second derivatives are set to zero, and a Newton update can be done efficiently, reducing to row-wise

optimization of U, V, and Z. For the subclass of models where Equations 6-8 are differentiable and the loss is decomposable define,

$$q(U_{i\cdot}) = \alpha \left(W_{i\cdot} \odot \left(f_1(U_{i\cdot}V^T) - X_{i\cdot}\right)\right) V + A_{i\cdot}$$
$$q(V_{i\cdot}) = \alpha \left(W_{i\cdot} \odot \left(f_1(UV_{i\cdot}^T) - X_{i\cdot}\right)\right)^T U + (1-\alpha)\left(\tilde{W}_{i\cdot} \odot \left(f_2(V_{i\cdot}Z^T) - Y_{i\cdot}\right)\right) Z + B_{i\cdot}$$
$$q(Z_{i\cdot}) = (1-\alpha)\left(\tilde{W}_{\cdot i} \odot \left(f_2(VZ_{i\cdot}^T) - Y_{\cdot i}\right)\right)^T V + C_{i\cdot}$$

Any local optimum of the loss corresponds to roots of the equations $\{q(U_{i\cdot})\}_{i=1}^m$, $\{q(V_{i\cdot})\}_{i=1}^n$, and $\{q(Z_{i\cdot})\}_{i=1}^r$. Using a Newton step, the update for $U_{i\cdot}$ is

$$U_{i\cdot}^{\text{new}} = U_{i\cdot} - \eta \left[q(U_{i\cdot})\right]\left[q'(U_{i\cdot})\right]^{-1}, \tag{9}$$

where $\eta \in [0,1]$ is the step length, chosen using line search with the Armijo criterion [3, ch. 3]. The Newton steps for $V_{i\cdot}$ and $Z_{i\cdot}$ are analogous. To describe the Hessians of the loss, q', we introduce the following notation for the Hessians of G^*, H^* and I^*:

$$G_i \equiv \text{diag}(\nabla^2 G^*(U_{i\cdot})), \quad H_i \equiv \text{diag}(\nabla^2 H^*(V_{i\cdot})), \quad I_i \equiv \text{diag}(\nabla^2 I^*(Z_{i\cdot})).$$

For case of ℓ_2-regularization $G_i = \text{diag}(a^{-1}\mathbf{1})$. For conciseness we also introduce the following terms:

$$D_{1,i} \equiv \text{diag}(W_{i\cdot} \odot f_1'(U_{i\cdot}V^T)), \quad D_{2,i} \equiv \text{diag}(W_{\cdot i} \odot f_1'(UV_{i\cdot}^T)),$$
$$D_{3,i} \equiv \text{diag}(\tilde{W}_{i\cdot} \odot f_2'(V_{i\cdot}^T Z)), \quad D_{4,i} \equiv \text{diag}(\tilde{W}_{\cdot i} \odot f_2'(VZ_{i\cdot}^T)).$$

This allows us to describe the Hessians of Equation 1 with respect to each row of the parameter matrices.

Lemma 4. *The Hessians of Equation 1 with respect to $U_{i\cdot}$, $V_{i\cdot}$, and $Z_{i\cdot}$ are*

$$q'(U_{i\cdot}) \equiv \frac{\partial q(U_{i\cdot})}{\partial U_{i\cdot}} = \alpha V^T D_{1,i} V + G_i$$
$$q'(Z_{i\cdot}) \equiv \frac{\partial q(Z_{i\cdot})}{\partial Z_{i\cdot}} = (1-\alpha) V^T D_{4,i} V + I_i$$
$$q'(V_{i\cdot}) \equiv \frac{\partial q(V_{i\cdot})}{\partial V_{i\cdot}} = \alpha U^T D_{2,i} U + (1-\alpha) Z^T D_{3,i} Z + H_i$$

Proof. We prove the result for $q'(U_{i\cdot})$, noting that the other derivations are similar. Since $q(\cdot)$ and its argument $U_{i\cdot}$ are both vectors, $Dq(U_{i\cdot})$ is identical to $\partial q(U_{i\cdot})/\partial U_{i\cdot}$. Ignoring the terms that do not vary with $U_{i\cdot}$,

$$\begin{aligned}
Dq(U_{i\cdot}) &= D\left[\alpha(W_{i\cdot} \odot f(U_{i\cdot}V^T))V + A_{i\cdot}\right] \\
&= \alpha D\left[(W_{i\cdot} \odot f(U_{i\cdot}V^T))V\right] + DA_{i\cdot} \\
&= \alpha\left\{(V^T \otimes I_1) \times \frac{\partial \,\text{vec}(W_{i\cdot} \odot f(U_{i\cdot}V^T))}{\partial \,\text{vec}\, U_{i\cdot}} + (I_k \otimes V)\frac{\partial \,\text{vec}\, V}{\partial \,\text{vec}\, U_{i\cdot}}\right\} + \frac{\partial \,\text{vec}\, \nabla^2 G^*(U_{i\cdot})}{\partial \,\text{vec}\, U_{i\cdot}} \\
&= \alpha\left\{(V^T \otimes I_1) \times D_{1,i}V + (I_k \otimes V) \times 0\right\} + G_i \\
&= \alpha V^T D_{1,i} V + G_i.
\end{aligned}$$

□

Plugging the gradient $q(U_{i\cdot})$ and the Hessian $q'(U_{i\cdot})$ into Equation 9 yields

$$\begin{aligned}
U_{i\cdot}^{\text{new}} q'(U_{i\cdot}) &= U_{i\cdot}(\alpha V^T D_{1,i} V + G_i) + \alpha\eta\left(W_{i\cdot} \odot \left(X_{i\cdot} - f(U_{i\cdot}V^T)\right)\right) V - \eta A_{i\cdot} \\
&= \alpha U_{i\cdot} V^T D_{1,i} V + \alpha\eta\left(W_{i\cdot} \odot \left(X_{i\cdot} - f(U_{i\cdot}V^T)\right)\right) D_{1,i}^{-1} D_{1,i} V + U_{i\cdot} G_i - \eta A_{i\cdot} \\
&= \alpha\left(U_{i\cdot}V^T + \eta\left(W_{i\cdot} \odot \left(X_{i\cdot} - f(U_{i\cdot}V^T)\right) D_{1,i}^{-1}\right) D_{1,i} V + U_{i\cdot} G_i - \eta A_{i\cdot}\right.
\end{aligned} \tag{10}$$

Likewise for $Z_{i\cdot}$,

$$Z_{i\cdot}^{\text{new}} q'(Z_{i\cdot}) = (1-\alpha)\left(Z_{i\cdot}V^T + \eta\left(\tilde{W}_{\cdot i} \odot \left(Y_{\cdot i} - f(VZ_{i\cdot}^T)\right)\right)^T D_{4,i}^{-1}\right) D_{4,i}V + Z_{i\cdot}I_i - \eta C_{i\cdot} \tag{11}$$

The derivation of the update for $V_{i\cdot}$ is similar, since $L(U, V, Z|W, \tilde{W})$ is a linear combination and the differential operator is linear:

$$\begin{aligned} V_{i\cdot}^{\text{new}} q'(V_{i\cdot}) = \alpha &\left\{\left(V_{i\cdot}U^T + \eta\left(W_{\cdot i} \odot \left(X_{\cdot i} - f(UV_{i\cdot}^T)\right)\right)^T D_{2.i}^{-1}\right) D_{2,i}U\right\} + \\ (1-\alpha) &\left\{\left(V_{i\cdot}Z^T + \eta\left(\tilde{W}_{i\cdot} \odot \left(Y_{i\cdot} - f(V_{i\cdot}Z^T)\right)\right) D_{3,i}^{-1}\right) D_{3,i}Z\right\} + \\ &V_{i\cdot}H_i - \eta B_{i\cdot} \end{aligned} \tag{12}$$

While $D \in \{D_{j,i}\}_{j=1}^4$ may not be invertible, *i.e.*, a diagonal entry is zero when the corresponding weight is zero, the form of the update equations shows that this does not matter. If a diagonal entry in D is zero, replacing its corresponding entry in D^{-1} by any nonzero value does not change the result of Equations 10, 11, and 12, as the zero weight cancels it out.

References

[1] J. R. Magnus and K. M. Abadir. On some definitions in matrix algebra. Working Paper CIRJE-F-426, University of Tokyo, Feb. 2007.

[2] J. R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley, 2007.

[3] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research. Springer, 1999.