Past and new developments in pairwise likelihood estimation for latent variable models

Irini Moustaki

collaborators: Karl Jöreskog, Myrsini Katsikatsou, Silvia Cagnone, Vassilis Vasdekis, Dimitris Rizopoulos, Ioulia Papagheorgiou, Chris Skinner, Haziq Jamil, Yunxiao Chen, Guiseppe Alfonzeti, Ruggero Bellio

Outline

- Brief introduction to latent variable models
- Observed variables: binary, ordinal, and continuous
- Modeling: Structural Equation Modeling (SEM)
- Methodology discussed: Pairwise Likelihood (PL)
- Topics that will be discussed:
    - Estimation
    - Overall goodness-of-fit testing: nested models and overall fit under SRS
    - Limited goodness-of-fit tests under SRS and complex sample designs
    - Model selection criteria
    - Reducing computational complexity

Latent variables and measurement

Using statistical models to understand constructs better: a question of **measurement**

- Many theories in behavioral and social sciences are formulated in terms of theoretical constructs that are not directly observed

  attitudes, opinions, abilities, motivations, etc.

- The measurement of a construct is achieved through one or more observable **indicators** (questionnaire **items**, tests).
- The purpose of a measurement model is to describe how well the observed indicators serve as a measurement instrument for the constructs, also known as **latent variables**.
- **Measurement models** often suggest ways in which the observed measurements can be improved.

Latent variables and substantive theories

Using statistical models to understand relationships between constructs and covariates and to test **theories** about those relationships.

- Often measurement by multiple indicators may involve more than one latent variable.
- Subject-matter theories and research questions usually concern relationships among the latent variables, and perhaps also observed explanatory variables.
- Latent variables can be used as predictors for distal outcomes or as dependent variables explained by covariates.
- These are captured by statistical models for those variables: **structural models**.

Motivation of our work

- Improve the estimation in cases of intractable integrals and complex models.
- Provide an inferential framework for model testing and model selection.
- Improve the computational time and cost.

Notation

- $\mathbf{y}$: $p-$dimensional vector of the observed variables (binary, ordinal, continuous).
- $\mathbf{y}^{\star}$: $p-$dimensional vector of corresponding underlying continuous variables.
- The connection between y_i and $y_i^{\star}$ is

$$y_i = c_i \Longleftrightarrow \tau_{c_i - 1}^{(y_i)} < y_i^{\star} < \tau_{c_i}^{(y_i)}, \tag{1}$$

$$-\infty = \tau_0^{(y_i)} < \tau_1^{(y_i)} < \ldots < \tau_{m_i - 1}^{(y_i)} < \tau_{m_i}^{(y_i)} = +\infty.$$

- c : the c-th response category of variable y_i, $c = 1, \ldots, m_i$, $\tau_{i,c}$: the c-th threshold of variable y_i,
- In practice, $y_i^{\star} \sim N(0, 1)$
- y_i is continuous: $y_i = y_i^{\star}$.

Structural Equation Model

Following Muthén (1984):

$$\mathbf{y}^{\star} = \boldsymbol{\nu} + \Lambda\boldsymbol{\eta} + \boldsymbol{\epsilon}$$
$$\boldsymbol{\eta} = \boldsymbol{\alpha} + \mathrm{B}\boldsymbol{\eta} + \Gamma\mathbf{x} + \boldsymbol{\zeta}$$

$\boldsymbol{\eta}$: vector of latent variables, $q-$dimensional,
$\mathbf{x}$: vector of covariates,
$\boldsymbol{\epsilon}$ and $\boldsymbol{\zeta}$: vectors of error terms, and
$\boldsymbol{\nu}$ and $\boldsymbol{\alpha}$: vectors of intercepts.
Standard assumptions:

- $\boldsymbol{\eta}$, $\boldsymbol{\epsilon}$, $\boldsymbol{\zeta}$ follow multivariate normal distribution,
- $Cov\left(\boldsymbol{\eta}, \boldsymbol{\epsilon}\right) = Cov\left(\boldsymbol{\eta}, \boldsymbol{\zeta}\right) = Cov\left(\boldsymbol{\epsilon}, \boldsymbol{\zeta}\right) = \mathbf{0}$,
- $I - \mathrm{B}$ is non-singular, I the identity matrix.

Structural Equation Model

Based on the model:

$$\boldsymbol{\mu} \equiv E\left(\mathbf{y}^{\star}|\mathbf{x}\right) = \boldsymbol{\nu} + \Lambda\left(I - \mathrm{B}\right)^{-1}\left(\boldsymbol{\alpha} + \Gamma\mathbf{x}\right)$$

$$\Sigma \equiv Cov\left(\mathbf{y}^{\star}|\mathbf{x}\right) = \Lambda\left(I - \mathrm{B}\right)^{-1}\Psi\left[\left(I - \mathrm{B}\right)^{-1}\right]'\Lambda' + \Theta$$

Let $\boldsymbol{\theta}$ be the parameter vector of the model.

$$\boldsymbol{\theta}' = \left(\mathrm{vec}\left(\Lambda\right)', \mathrm{vec}\left(B\right)', \mathrm{vec}\left(\Gamma\right)', \mathrm{vech}\left(\Psi\right)', \mathrm{vech}\left(\Theta\right)', \boldsymbol{\alpha}', \boldsymbol{\nu}', \boldsymbol{\tau}'\right)$$

.

Likelihood Function

- Under the model, the probability of a response pattern r is

$$\pi_r(\boldsymbol{\theta}) = \pi\left(y_1 = c_1, \ldots, y_p = c_p; \boldsymbol{\theta}\right) = \int \ldots \int \phi_p(\mathbf{y}^\star; \Sigma_{\mathbf{y}^\star}) d\mathbf{y}^\star \ , \quad (2)$$

  where $\phi_p(\mathbf{y}^\star; \Sigma_{\mathbf{y}^\star})$ is a p-dimensional normal density with zero mean, and correlation matrix $\Sigma_{\mathbf{y}^\star}$.
- The maximization of log-likelihood over the parameter vector $\boldsymbol{\theta}$ requires the evaluation of the p-dimensional integral which cannot be written in a closed form.
- Maximum likelihood infeasible for large number of observed variables.

Alternative estimation: WLS and Composite Methods

- Three-stage estimation methods (Jöreskog, 1990, 1994; Muthén, 1984): unweighted least squares (ULS), diagonally weighted least squares (DWLS), and weighted least squares (WLS).

- Composite likelihood estimation (Besag 1974; Lindsay 1988; Cox and Reid 2004; Varin, Reid and Firth 2011).
  - Pairwise likelihood estimation for SEM (Jöreskog and Moustaki 2001; Katsikatsou, et al. 2012).

Pairwise likelihood estimation

Denote by $\{A_1, \cdots, A_K\}$ a set of conditional or marginal events with associated likelihoods $L_k(\boldsymbol{\theta}; \mathbf{y})$.
Following Lindsay (1988) a composite likelihood is the weighted product

$$L_k(\boldsymbol{\theta}; \mathbf{y}) = \prod_{k=1}^{K} L_k(\boldsymbol{\theta}; \mathbf{y})^{w_k},$$

where w_k are non-negative weights.

Following Cox & Reid (2004), the composite-loglikelihood could be modified as follows:

$$l(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i<j} \ln L\left(\boldsymbol{\theta}; (y_i, y_j)\right) - c \sum_{i} \ln L\left(\boldsymbol{\theta}; y_i\right) \ ,$$

where c is a constant to be chosen for optimal efficiency.

Pairwise likelihood for SEM

Basic assumption:

$$\begin{pmatrix} y_i^\star \\ y_j^\star \end{pmatrix} \Bigg| \mathbf{x} \sim N_2 \left(\begin{pmatrix} \mu_i \\ \mu_j \end{pmatrix}, \begin{pmatrix} \sigma_{ii} & \\ \sigma_{ji} & \sigma_{jj} \end{pmatrix} \right)$$

The *pl* for N independent observations:

$$pl\,(\boldsymbol{\theta}; \mathbf{y}|\mathbf{x}) = \sum_{n=1}^{N} \sum_{i<j} \ln L\,(\boldsymbol{\theta}; (y_{in}, y_{jn})|\mathbf{x}) \ .$$

The specific form of $\ln L\,(\boldsymbol{\theta}; (y_{in}, y_{jn})|\mathbf{x})$ depends on the type of the observed variables (binary/ ordinal, continuous).

Pairwise Likelihood Estimation for Binary Responses (1)

- For a pair of variables y_i and y_j. The basic pairwise log-likelihood takes the form

$$\sum_{i<j} \sum_{c_i=0}^{1} \sum_{c_j=0}^{1} n_{c_i c_j}^{(y_i y_j)} \ln \pi_{c_i c_j}^{(y_i y_j)}(\boldsymbol{\theta}) \tag{3}$$

where $n_{c_i c_j}$ is the observed frequency of sample units with $y_i = c_i$ and $y_j = c_j$.

- To accommodate complex sampling, the PL becomes:

$$pl(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i<j} \sum_{c_i=0}^{1} \sum_{c_j=0}^{1} p_{c_i c_j}^{(y_i y_j)} \ln \pi_{c_i c_j}^{(y_i y_j)}(\boldsymbol{\theta}) \ , \tag{4}$$

where $p_{c_i c_j} = \sum_{h \in s} w_h I(y_i^{(h)} = c_i, y_j^{(h)} = c_j) / \sum_{h \in s} w_h$.

Pairwise Likelihood Estimation for Binary Responses (2)

The score function

$$\nabla pl(\boldsymbol{\theta};\mathbf{y}) = \sum_{i<j}\sum_{c_i=0}^{1}\sum_{c_j=0}^{1} p_{c_i c_j}^{(y_i y_j)} (\pi_{c_i c_j}^{(y_i y_j)}(\boldsymbol{\theta}))^{-1} \frac{\partial \pi_{c_i c_j}^{(y_i y_j)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \qquad (5)$$

Using Taylor expansion, we may write

$$\hat{\boldsymbol{\theta}}_{PL} = \boldsymbol{\theta} + H(\boldsymbol{\theta})^{-1}\nabla pl(\boldsymbol{\theta};\mathbf{y}) + o_p(n^{-1/2}) \qquad (6)$$

where $H(\boldsymbol{\theta})$ is the sensitivity matrix, $H(\boldsymbol{\theta}) = E\left\{-\nabla^2 pl(\boldsymbol{\theta};\mathbf{y})\right\}$. It follows that

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{PL} - \boldsymbol{\theta}\right) \xrightarrow{d} N_t\left(0, H(\boldsymbol{\theta})J^{-1}(\boldsymbol{\theta})H(\boldsymbol{\theta})\right) ,$$

where t is the dimension of $\boldsymbol{\theta}$, and $J(\boldsymbol{\theta})$ is the variability matrix, $J(\boldsymbol{\theta}) = Var\left\{\sqrt{n}\nabla pl(\boldsymbol{\theta};\mathbf{y})\right\}$.

Why PL is proposed?

Maximum likelihood (ML) is not feasible for large models.
It requires the computation of multiple integrals over a multivariate normal distribution the dimension of which is equal to the number of ordinal observed variables.

Three stage least squares methods require the estimation of a weight matrix to obtain correct standard errors and chi-squared test statistics. A relatively large sample size is required for a reliable estimate.

The construction of model selection criteria of AIC and BIC type is not possible.

Finite-sample properties of PL estimation

For factor analysis models with ordinal data (Katsikatsou et al., 2012):

- PL estimates and standard errors present a close-to-zero bias and mean squared error (MSE).
- PL performs very similarly to three-stage least squares methods and maximum likelihood as implemented in the GLLVM approach.

Model fit

Katsikatsou and Moustaki, 2016.

- Pairwise Likelihood Ratio Test (PLRT) for overall fit
- Pairwise Likelihood Ratio Test for comparing models (e.g. equality constraints)
- Model selection criteria: PL versions of AIC and BIC
- The PLRT statistic performs in accordance with the asymptotic results at 5% and 1% significance levels for $N = 500, 1000$ but not satisfactorily for $N = 200$.
- Both adjusted AIC and BIC criteria perform very well with a minimum rate of success 82.9%.

Software

In the **R package** `lavaan`

PL available for fitting and testing factor analysis models or SEMs where

- all observed variables are binary or ordinal, and
- the standard parametrization for the underlying variables is used (zero means and unit variances)
- Multigroup analysis is also possible.

Current work

- Limited information test statistics under SRS and complex designs.
- Methods for reducing the computational complexity of pairwise estimation
  - Employ sampling methodology for selecting pairs (Papageorgiou and Moustaki, 2019)
  - Stochastic optimization

Fit on the Lower order margins

- Let $\dot{\boldsymbol{\pi}}_1 = (P(y_1 = 1), P(y_2 = 1), \ldots, P(y_p = 1))'$ be the $p \times 1$ vector that contains all univariate probabilities of a positive response to an item.
- Let $\dot{\boldsymbol{\pi}}_2$ be the $\binom{p}{2} \times 1$ vector of bivariate probabilities with elements, $\dot{\pi}_{ij} = P(y_i = 1, y_j = 1), j < i$.
- Let $\boldsymbol{\pi}_2$ be the vector that contains both these univariate and bivariate probabilities with dimension $s = p + \binom{p}{2} = p(p+1)/2$.
- We also define an $s \times 2^p$ indicator matrix T_2 of rank s such that $\boldsymbol{\pi}_2 = T_2 \boldsymbol{\pi}$.

Goodness-of-fit tests, simple hypothesis

- Let us denote with $\mathbf{p}$ the $2^p \times 1$ vector of sample proportions corresponding to the vector of population proportions $\boldsymbol{\pi}$. Assuming i.i.d, it is known that:

$$\sqrt{n}(\mathbf{p} - \boldsymbol{\pi}) \xrightarrow{d} N(0, \Sigma), \tag{7}$$

- where $\Sigma = D(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'$ and n is the sample size.
- Under complex sampling design, the vector $\mathbf{p}$ becomes the weighted vector of proportions $\mathbf{p}$ with elements $\sum_{h \in s} w_h I(\mathbf{y}^{(h)} = \mathbf{y}_r) / \sum_{h \in s} w_h$.
- Under suitable conditions (e.g. Fuller, 2009, sect. 1.3.2) we still have a central limit theorem, where the covariance matrix Σ need now not take a multinomial form.

Limited information goodness-of-fit tests

Reiser (1996, 2008), Bartholomew and Leung (2002), Maydey-Olivares and Joe (2005, 2006) Cagnone and Mignani (2007).

The test statistics developed are based on marginal distributions rather than on the whole response pattern.

- $H_o : \boldsymbol{\pi}_2 = \boldsymbol{\pi}_2(\boldsymbol{\theta})$ for some $\boldsymbol{\theta}$ versus $H_1 : \boldsymbol{\pi}_2 \neq \boldsymbol{\pi}_2(\boldsymbol{\theta})$ for any $\boldsymbol{\theta}$.
- Construct test statistics based upon the residual vector $\hat{\mathbf{e}}_2 = \mathbf{p}_2 - \boldsymbol{\pi}_2(\hat{\boldsymbol{\theta}}_{PL})$ derived from the bivariate marginal distributions of $\mathbf{y}$.
- We first derive the asymptotic distribution of $\hat{\mathbf{e}}_2$.

Limited information goodness-of-fit tests

- Following earlier notation, we can write $s \times 1$ vectors: $\boldsymbol{\pi}_2 = T_2\boldsymbol{\pi}$ and $\mathbf{p}_2 = T_2\mathbf{p}$. It follows that:

$$\sqrt{n}(\mathbf{p}_2 - \boldsymbol{\pi}_2) \xrightarrow{d} N(0, \Sigma_2), \tag{8}$$

  where $\Sigma_2 = T_2\Sigma T_2'$. Because T_2 is of full rank s, Σ_2 is also of full rank s.

Limited information goodness-of-fit tests

Noting that $\boldsymbol{\pi}_2(\boldsymbol{\theta}) = T_2\boldsymbol{\pi}(\boldsymbol{\theta})$, a Taylor series expansion gives:

$$\boldsymbol{\pi}_2(\hat{\boldsymbol{\theta}}_{PL}) = \boldsymbol{\pi}_2(\boldsymbol{\theta}) + T_2\Delta(\hat{\boldsymbol{\theta}}_{PL} - \boldsymbol{\theta}) + o_p(n^{-1/2}), \tag{9}$$

where $\Delta = \frac{\partial \boldsymbol{\pi}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$
Hence, using (6), we have

$$\hat{\mathbf{e}}_2 = \mathbf{p}_2 - \boldsymbol{\pi}_2(\boldsymbol{\theta}) - T_2\Delta H(\boldsymbol{\theta})^{-1}\nabla pl(\boldsymbol{\theta}; \mathbf{y}) + o_p(n^{-1/2}). \tag{10}$$

Finally we need to express $\nabla pl(\boldsymbol{\theta}; \mathbf{y})$ in terms of $\mathbf{p}_2 - \boldsymbol{\pi}_2(\boldsymbol{\theta})$

Limited information goodness-of-fit tests

Hence, there is a $t \times s$ matrix $B(\boldsymbol{\theta})$ such that

$$\nabla pl(\boldsymbol{\theta}; \mathbf{y}) = B(\boldsymbol{\theta})(\mathbf{p}_2 - \boldsymbol{\pi}_2(\boldsymbol{\theta})) \tag{11}$$

Hence, from (10)

$$\hat{\mathbf{e}}_2 = (I - T_2 \Delta H(\boldsymbol{\theta})^{-1} B(\boldsymbol{\theta}))(\mathbf{p}_2 - \boldsymbol{\pi}_2(\boldsymbol{\theta})) + o_p(n^{-1/2}) \tag{12}$$

So from (8), we have under H_0 that:

$$\sqrt{n}\hat{\mathbf{e}}_2 \xrightarrow{d} N(0, \Omega). \tag{13}$$

where $\Omega = (I - T_2 \Delta H(\boldsymbol{\theta})^{-1} B(\boldsymbol{\theta}))\Sigma_2 (I - T_2 \Delta H(\boldsymbol{\theta})^{-1} B(\boldsymbol{\theta}))'$.

Limited information goodness-of-fit tests

To estimate the asymptotic covariance matrix of $\hat{\mathbf{e}}_2$, we evaluate $\frac{\partial \boldsymbol{\pi}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ at the PL estimate $\hat{\boldsymbol{\theta}}_{PL}$ to obtain $\hat{\Delta}$ and set:

$$\hat{\Omega} = (I - T_2 \hat{\Delta} \hat{H}(\hat{\boldsymbol{\theta}}_{PL})^{-1} B(\hat{\boldsymbol{\theta}}_{PL})) \hat{\Sigma}_2 (I - T_2 \hat{\Delta} \hat{H}(\hat{\boldsymbol{\theta}}_{PL})^{-1} B(\hat{\boldsymbol{\theta}}_{PL}))',$$

where $\hat{\Sigma}_2 = T_2 \hat{\Sigma} T_2'$. In the case of iid observations with a multinomial covariance matrix, we may set $\hat{\Sigma} = D(\mathbf{p}) - \mathbf{p}\mathbf{p}'$.

Wald Test Statistic

- A Wald test statistic is given by:

$$L_2 = n(\mathbf{p}_2 - \boldsymbol{\pi}_2(\hat{\theta}_{PL}))'\hat{\Omega}^{+}(\mathbf{p_2} - \boldsymbol{\pi}_\mathbf{2}(\hat{\boldsymbol{\theta}}_\mathbf{PL})), \tag{14}$$

- where $\hat{\Omega}^{+}$ is the Moore-Penrose inverse of $\hat{\Omega}$.
- Under H_0, L_2 is asymptotically distributed as χ^2 with d.f. equal to the rank of $\hat{\Omega}^{+}$, which is between $s - t$ and s.

Pearson Chi-square Test Statistic

- Pearson test statistic, let D_2 be the $s \times s$ matrix $D_2 = diag(\pi_2(\boldsymbol{\theta}))$ and let $\hat{D}_2 = diag(\pi_2(\hat{\boldsymbol{\theta}}_{PL}))$. Then the Pearson test statistic is given by

$$X_P^2 = n\hat{\mathbf{e}}_2' \hat{D}_2^{-1} \hat{\mathbf{e}}_2 = n(\mathbf{p}_2 - \boldsymbol{\pi}_2(\hat{\boldsymbol{\theta}}_{PL}))' \hat{D}_2^{-1} (\mathbf{p_2} - \boldsymbol{\pi_2}(\hat{\boldsymbol{\theta}}_{\mathbf{PL}})). \qquad (15)$$

- The limiting distribution of $\sqrt{n}\hat{D}_2^{-0.5}\hat{\mathbf{e}}_2$ under the hypothesis that the model is correct is given by $N(0, {D_2}^{-0.5}\Omega {D_2}^{-0.5})$.
- Hence X_P^2 has the limiting distribution of $\sum \delta_i W_i$, where the δ_i are eigenvalues of ${D_2}^{-0.5}\Omega {D_2}^{-0.5}$ and the W_i are independent chi-square random variables, each with one degree of freedom.
- These eigenvalues can be estimated by the eigenvalues of $\hat{D}_2^{-0.5}\hat{\Omega}\hat{D}_2^{-0.5}$.
- A first and a second order Rao-Scott type test can be obtained.

Simulation results, SRS

Empirical Type I error probabilities for the Wald tests and FSMAdj Pearson tests, $N = 1000$.

Simulation Study	Wald 5%	Wald 1%	FSMAdj Pearson 5%	FSMAdj Pearson 1%
1F 5Items	0.053	0.009	0.050	0.012
1F 8Items	0.055	0.011	0.051	0.010
1F 10Items	0.059	0.038	0.078	0.022
2F 10Items	0.059	0.017	0.059	0.016
3F 15Items	0.023	0.011	0.072	0.023

Estimation of the covariance matrix under complex sampling: stratified multistage sampling

- $$\begin{aligned} \Sigma &= limvar\{\sqrt{n}(\mathbf{p} - \boldsymbol{\pi})\} \\ &= limvar\{\sqrt{n}(\frac{\sum_{h\in s} w_h \mathbf{y}^{(h)}}{\sum_{h\in s} w_h} - \boldsymbol{\pi})\} \end{aligned}$$

  where *limvar* denotes the asymptotic covariance matrix.
- Using a usual linearization argument for a ratio:

$$\Sigma = limvar\{\sqrt{n}\frac{\sum_{h\in s} w_h(\mathbf{y}^{(h)} - \boldsymbol{\pi})}{E(\sum_{h\in s} w_h)}\}. \tag{16}$$

Current research

- Study the performance of the Wald and Pearson chi-square test statistics under different simulation senaria under SRS and complex survey designs in terms of Type I error and power.
- Implement the tests in real data sets.

Why use Stochastic Optimization? (Yunxiao Chen, Guiseppe Alfonzeti, Ruggero Bellio)

Pros and cons of pairwise likelihood:

+ It substitutes large-dimensional integration problems with bivariate ones.
- Its computational cost grows with the number of pairs, $O(p^2)$.

Using stochastic optimization:

- Resampling a new small subset of the data at each iteration
- Low computational cost per iteration and low memory storage
- In our case $p\ell(\boldsymbol{\theta}; \mathbf{y})$ depends on the data only trough the bivariate frequencies $n^{ij}_{s_i s_j}$, such that sampling units across iterations does not reduce complexity.
- Reducing the number of pairs is proposed here.

Overview of Stochastic Optimization

- Define a stochastic approximation to $p\ell(\boldsymbol{\theta}, \mathbf{y})$ via

$$f(\boldsymbol{\theta}; \mathbf{y}, \mathbf{w}) \propto \sum_{i<j} w_{ij} \ell_{ij}(\boldsymbol{\theta});$$

- The quantities w_{ij} are random binary weights such that

$$w_{ij} \overset{iid}{\sim} \text{Bernoulli}(\gamma);$$

- The hyperparameter $\gamma \in (0, 1]$ controls the trade-off between the accuracy of the approximation and its computational complexity.

- The complexity of $f(\boldsymbol{\theta}; \mathbf{y}, \mathbf{w})$ grows with $O(\gamma p^2)$. It follows that, if γ is set at the same order of p^{-1}, the complexity of the approximation grows only linearly in p.

Stochastic Optimization - The algorithm

The generic t-th iteration is performed alternating:

1. **Stochastic step:**
   - Sample a new set of weights $\mathbf{w}^{(t)}$;
2. **Approximation step:**
   - Build a cheap approximation of

$$\nabla f(\boldsymbol{\theta}_{t-1}; \mathbf{y}, \mathbf{w}^{(t)}) = \frac{1}{\gamma} \sum_{i<j} w_{ij}^{(t)} \nabla \ell_{ij}(\theta_{t-1}; \mathbf{y});$$

   Note that, if $\gamma = 1$, we retrieve $\nabla f(\boldsymbol{\theta}_{t-1}; \mathbf{y}, \mathbf{w}^{(t)}) = \nabla p\ell(\boldsymbol{\theta}_{t-1})$. If $\gamma \neq 1$ we still have $E_w[\nabla f(\boldsymbol{\theta}_{t-1}; y, \mathbf{w}^{(t)})] = \nabla p\ell(\boldsymbol{\theta}_{t-1})$.
3. **Update step:**
   - Update θ_t via $\boldsymbol{\theta}_t = \text{Proj}_\Theta \left(\boldsymbol{\theta}_{t-1} + \eta_t \nabla f(\boldsymbol{\theta}_{t-1}; \mathbf{y}, \mathbf{w}^{(t)})\right)$, where $\text{Proj}_\Theta(.)$ is a projection operator which ensures $\rho_{ij}^{y^*}$ to be valid correlations. The stepsize used is $\eta_t = t^{-.5+\epsilon}$, with ϵ a small positive constant such that $\sum_{t=1}^{\infty} \eta_t = \infty$ and $\sum_{t=1}^{\infty} \eta_t^2 < \infty$, as in Zhang and Chen (2020).