Theorem 1: Attention weight properties

Let $\mathbf{x}_i \in \mathbb{R}^p$ represent individual response patterns, and let $\mathbf{a}_i(\mathbf{x}_i;\theta) \in \mathbb{R}^p$ denote the attention weights computed by the CFASA model with parameters $\theta \in \Theta$. Under Regularity Conditions (R1)–(R3), the attention weights possess the following fundamental properties:

(P1) *Probability Simplex Property*: For all i and all $\theta \in \Theta$:

$$\mathbf{a}_i \in \Delta^{p-1} := \left\{\mathbf{w} \in \mathbb{R}^p : w_j \geq 0, \sum_{j=1}^{p} w_j = 1\right\}$$

(P2) *Continuous Differentiability*: The mapping $(\mathbf{x}_i, \theta) \mapsto \mathbf{a}_i(\mathbf{x}_i;\theta)$ is continuously differentiable with respect to all arguments.

(P3) *Uniform Lipschitz Continuity*: There exists a constant $L > 0$ such that for all $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$ and all $\theta \in \Theta$:

$$\| \mathbf{a}_i(\mathbf{x}_i;\theta) - \mathbf{a}_j(\mathbf{x}_j;\theta) \|_2 \leq L \| \mathbf{x}_i - \mathbf{x}_j \|_2$$

Regularity conditions

(R1) *Bounded Input Space*: The input space is compact: $\mathcal{X} \subset \mathbb{R}^p$ is a compact set.

(R2) *Parameter Boundedness*: The parameter space Θ is compact with:

$\| \mathbf{W}_\ell \|_F \leq M_W$ for all weight matrices
$\| \mathbf{b}_\ell \|_2 \leq M_b$ for all bias vectors
$\tau \in [\tau_{\min}, \tau_{\max}]$ where $0 < \tau_{\min} < \tau_{\max} < \infty$

(R3) *Encoder Architecture*: The encoder function $f_{\text{encoder}}(\mathbf{x};\theta_{\text{enc}})$ uses ReLU activations and satisfies standard neural network regularity conditions.

Proof.
Proof of Property (P1): Probability Simplex
The attention weights are computed through the softmax transformation:

$$a_{ij} = \frac{\exp(r_{ij}/\tau)}{\sum_{k=1}^{p} \exp(r_{ik}/\tau)}$$

where $\mathbf{r}_i = \mathbf{W}_3\mathbf{c}_i + \mathbf{b}_3$ are the raw attention logits.

Non-negativity: Since $\exp(\cdot) > 0$ for all real arguments, we have $a_{ij} > 0$ for all i, j.

Normalization: Direct computation shows:

$$\sum_{j=1}^{p} a_{ij} = \sum_{j=1}^{p} \frac{\exp(r_{ij}/\tau)}{\sum_{k=1}^{p} \exp(r_{ik}/\tau)}$$

$$= \frac{1}{\sum_{k=1}^{p} \exp(r_{ik}/\tau)} \sum_{j=1}^{p} \exp(r_{ij}/\tau)$$

$$= \frac{\sum_{j=1}^{p} \exp(r_{ij}/\tau)}{\sum_{k=1}^{p} \exp(r_{ik}/\tau)} = 1$$

Therefore, $\mathbf{a}_i \in \Delta^{p-1}$ for all i and θ.

Proof of Property (P2): Continuous Differentiability
We establish differentiability by examining each component of the composition:

$$\mathbf{a}_i = \text{softmax}\left(\frac{\mathbf{W}_3 f_{\text{encoder}}(\mathbf{x}_i;\theta_{\text{enc}})+\mathbf{b}_3}{\tau}\right)$$

Step 1: Encoder Differentiability. The ReLU function $\sigma(t) = \max(0, t)$ is differentiable almost everywhere with derivative:

$$\sigma'(t) = \begin{cases} 1 & \text{if} t > 0 \\ 0 & \text{if} t < 0 \end{cases}$$

For the two-layer encoder:

$$f_{\text{encoder}}(\mathbf{x}; \theta_{\text{enc}}) = \sigma(\mathbf{W}_2\sigma(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2)$$

The composition of ReLU networks is differentiable almost everywhere, and the set of non-differentiable points has measure zero.

Step 2: Linear Layer Differentiability. The raw logits $\mathbf{r}_i = \mathbf{W}_3\mathbf{c}_i + \mathbf{b}_3$ are linear in both the context vector $\mathbf{c}_i$ and the parameters $(\mathbf{W}_3, \mathbf{b}_3)$, hence continuously differentiable.

Step 3: Softmax Differentiability. The softmax function $\mathbf{s} = \text{softmax}(\mathbf{u})$ has Jacobian:

$$\frac{\partial s_j}{\partial u_k} = s_j(\delta_{jk} - s_k)$$

where δ_{jk} is the Kronecker delta. Since $s_j > 0$ for all j (from Property P1), this Jacobian is well-defined and continuous.

Step 4: Temperature Differentiability. For the temperature parameter:

$$\frac{\partial a_{ij}}{\partial \tau} = -\frac{1}{\tau^2}\sum_{k=1}^{p}\frac{\partial s_j}{\partial u_k} r_{ik}$$

Under condition (R2), $\tau > \tau_{\min} > 0$, ensuring the derivative is well-defined.
Chain Rule Application: By the chain rule and the continuous differentiability of each component, the entire mapping is continuously differentiable in $(\mathbf{x}_i, \theta)$.

Proof of Property (P3): Uniform Lipschitz Continuity
We establish Lipschitz continuity through a series of lemmas.
Under condition (R2), the encoder function satisfies:

$$\| f_{\text{encoder}}(\mathbf{x}; \theta) - f_{\text{encoder}}(\mathbf{x}'; \theta) \|_2 \leq L_{\text{enc}} \| \mathbf{x} - \mathbf{x}' \|_2$$

Proof of Lemma. For ReLU networks, the Lipschitz constant is bounded by the product of spectral norms:

$$L_{\text{enc}} = \| \mathbf{W}_2 \|_2 \cdot \| \mathbf{W}_1 \|_2 \leq M_W^2$$

This follows from the fact that ReLU has Lipschitz constant 1, and compositions of Lipschitz functions have Lipschitz constants bounded by the product.

The raw logits satisfy:

$$\| \mathbf{r}_i - \mathbf{r}_j \|_2 \leq \| \mathbf{W}_3 \|_2 \cdot \| \mathbf{c}_i - \mathbf{c}_j \|_2 \leq M_W L_{\text{enc}} \| \mathbf{x}_i - \mathbf{x}_j \|_2$$

For the temperature-scaled softmax:

$$\| \text{softmax}(\mathbf{u}/\tau) - \text{softmax}(\mathbf{v}/\tau) \|_2 \leq \frac{1}{\tau} \| \mathbf{u} - \mathbf{v} \|_2$$

Proof of Lemma. The softmax function has Lipschitz constant 1 when applied to vectors with bounded norm. The temperature scaling introduces the factor $1/\tau$, and condition (R2) ensures $\tau \geq \tau_{\min} > 0$.

Main Proof Completion: Combining the lemmas:

$$\| \mathbf{a}_i - \mathbf{a}_j \|_2 \leq \frac{1}{\tau_{\min}} \| \mathbf{r}_i - \mathbf{r}_j \|_2$$

$$\leq \frac{M_W L_{\text{enc}}}{\tau_{\min}} \| \mathbf{x}_i - \mathbf{x}_j \|_2$$

$$= L \| \mathbf{x}_i - \mathbf{x}_j \|_2$$

where $L = \frac{M_W^3}{\tau_{\min}}$ is the uniform Lipschitz constant.

Theorem 2: CFASA-CFA equivalence

Let $\{(\mathbf{x}_i, \eta_i^*)\}_{i=1}^N$ be observations from the CFASA population model with parameters $\theta^* \in \Theta$. Under Equivalence Conditions (E1)–(E2), the CFASA model reduces to weighted confirmatory factor analysis with the following equivalence relationships:

(2a) *Parameter Equivalence*: When attention weights are uniform across individuals, there exists a weight vector $\mathbf{w} \in \Delta^{p-1}$ such that:

$$\mathbf{a}_i(\mathbf{x}_i; \theta^*) = \mathbf{w} \quad \forall i \in \{1,2,\dots,N\}$$

(2b) *Factor Score Equivalence*: The CFASA factor scores reduce to weighted linear combinations:

$$\hat{\eta}_i^{\text{CFASA}} = \mathbf{w}^T \mathbf{x}_i = \sum_{j=1}^p w_j x_{ij}$$

(2c) *Likelihood Equivalence*: The CFASA likelihood function reduces to the weighted CFA likelihood:

$$\mathcal{L}^{\text{CFASA}}(\theta^* | X) = \mathcal{L}^{\text{WCFA}}(\mathbf{w}, \sigma^2 | X)$$

where $\mathcal{L}^{\text{WCFA}}$ denotes the weighted confirmatory factor analysis likelihood with loading weights $\mathbf{w}$ and error variance σ^2.

Equivalence conditions

(E1) *Attention Invariance*: The attention weights are constant across individuals:

$$\mathbf{a}_i(\mathbf{x}_i; \theta) = \mathbf{a}(\theta) \quad \forall i$$

This occurs when either:
Response patterns are uninformative: $\frac{\partial \mathbf{a}(\mathbf{x};\theta)}{\partial \mathbf{x}} = \mathbf{0}$
Individual differences are absent: $\text{Var}[\mathbf{a}(X; \theta)] = \mathbf{0}$

(E2) *Regularity*: Standard factor analysis regularity conditions hold:
$E[\eta^*] = 0,\ \text{Var}[\eta^*] = 1$ (factor standardization)
$E[\boldsymbol{\varepsilon}] = \mathbf{0},\ \text{Cov}[\boldsymbol{\varepsilon}] = \sigma^2 \mathbf{I}_p$ (error assumptions)

Proof. The proof establishes equivalence at three levels: parameter structure, likelihood function, and distributional properties.

Step 1: Attention Weight Reduction
Under condition (E1), the attention weights become constant:

$$\mathbf{a}_i(\mathbf{x}_i; \theta) = \mathbf{w} \in \Delta^{p-1}$$

Case 1a (Uninformative Responses): If $\frac{\partial \mathbf{a}}{\partial \mathbf{x}} = \mathbf{0}$, then $\mathbf{a}$ reduces to a constant vector $\mathbf{w}$.
Case 1b (Homogeneous Population): If $\text{Var}[\mathbf{a}(X)] = \mathbf{0}$, then by definition:

$$\mathbf{a}(\mathbf{x}_i; \theta) = E[\mathbf{a}(X; \theta)] = \mathbf{w}$$

Step 2: Factor Score Equivalence
Given uniform attention weights $\mathbf{w}$, the CFASA factor score computation becomes:

$$\hat{\eta}_i^{\text{CFASA}} = \mathbf{a}_i^T \mathbf{x}_i$$

$$= \mathbf{w}^T \mathbf{x}_i$$

$$= \sum_{j=1}^{p} w_j x_{ij}$$

This is identical to the weighted factor score in traditional factor analysis with predetermined weights $\mathbf{w}$.
Comparison with Standard CFA: In traditional CFA, factor scores are computed as:

$$\hat{\eta}_i^{\text{CFA}} = \boldsymbol{\lambda}^T (\boldsymbol{\lambda}\boldsymbol{\lambda}^T + \boldsymbol{\Theta})^{-1} \mathbf{x}_i$$

When $\boldsymbol{\lambda} = c\mathbf{w}$ for some constant $c > 0$ and $\boldsymbol{\Theta} = \sigma^2 \mathbf{I}$, we get:

$$\hat{\eta}_i^{\text{CFA}} = \frac{c^2}{c^2+\sigma^2} \mathbf{w}^T \mathbf{x}_i$$

The CFASA and CFA factor scores are proportional with proportionality constant $\frac{c^2+\sigma^2}{c^2}$.

Step 3: Measurement Model Equivalence
Under conditions (E1)–(E2), the CFASA measurement model:

$$\mathbf{x}_i = \mathbf{a}_i(\mathbf{x}_i; \theta) \odot (\eta_i \mathbf{1}_p) + \boldsymbol{\varepsilon}_i$$

where $\odot$ denotes element-wise multiplication, reduces to:

$$\mathbf{x}_i = \mathbf{w} \odot (\eta_i \mathbf{1}_p) + \boldsymbol{\varepsilon}_i = \eta_i \mathbf{w} + \boldsymbol{\varepsilon}_i$$

This is equivalent to the weighted CFA measurement model:

$$\mathbf{x}_i = \boldsymbol{\lambda}\eta_i + \boldsymbol{\varepsilon}_i$$

where $\boldsymbol{\lambda} = \mathbf{w}$ (up to scaling).

Step 4: Likelihood Function Equivalence
CFASA Likelihood: Under the normality assumption, the CFASA likelihood is:

$$\mathcal{L}^{\text{CFASA}}(\theta|X) = \prod_{i=1}^{N} p(\mathbf{x}_i|\eta_i; \theta)$$

where:

$$p(\mathbf{x}_i|\eta_i; \theta) = \mathcal{N}(\mathbf{x}_i; \eta_i \mathbf{a}_i(\mathbf{x}_i; \theta), \sigma^2 \mathbf{I}_p)$$

Reduction under Uniform Attention: When $\mathbf{a}_i = \mathbf{w}$ for all i:

$$p(\mathbf{x}_i|\eta_i; \theta) = \mathcal{N}(\mathbf{x}_i; \eta_i \mathbf{w}, \sigma^2 \mathbf{I}_p)$$

Marginalization over Factors: Integrating over $\eta_i \sim \mathcal{N}(0,1)$:

$$p(\mathbf{x}_i; \theta) = \int p(\mathbf{x}_i|\eta_i; \theta) p(\eta_i)\, d\eta_i$$
$$= \mathcal{N}(\mathbf{x}_i; \mathbf{0}, \mathbf{w}\mathbf{w}^T + \sigma^2 \mathbf{I}_p)$$

Weighted CFA Likelihood: The weighted CFA likelihood with loadings $\boldsymbol{\lambda} = \mathbf{w}$ is:

$$\mathcal{L}^{\text{WCFA}}(\mathbf{w}, \sigma^2|X) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{x}_i; \mathbf{0}, \mathbf{w}\mathbf{w}^T + \sigma^2 \mathbf{I}_p)$$

Therefore: $\mathcal{L}^{\text{CFASA}} = \mathcal{L}^{\text{WCFA}}$.

Step 5: Parameter Space Embedding
The weighted CFA parameter space is embedded in the CFASA parameter space.

Define the embedding map $\phi: (\mathbf{w}, \sigma^2) \mapsto \theta$ where:

Encoder parameters are set to produce constant output: $f_{\text{encoder}}(\mathbf{x}; \theta_{\text{enc}}) = \mathbf{c}_0$
Loading predictor: $\mathbf{W}_3\mathbf{c}_0 + \mathbf{b}_3 = \tau \cdot \text{logit}(\mathbf{w})$
Error variance: σ^2 remains unchanged

where $\text{logit}(\mathbf{w}) = [\log(w_1/w_p), \ldots, \log(w_{p-1}/w_p)]^T$ is the inverse softmax transformation.

This embedding is:
Injective: Different $(\mathbf{w}, \sigma^2)$ map to different θ
Continuous: Small changes in $(\mathbf{w}, \sigma^2)$ produce small changes in θ
Measure-preserving: Likelihood functions are identical under the mapping

Implications

Performance Guarantee. CFASA performance is at least as good as weighted CFA since the latter is a special case.

Parameter Interpretation. When individual differences are absent, CFASA attention weights can be interpreted directly as factor loadings.

Model Selection. Standard CFA fit indices (CFI, RMSEA) can be applied to assess CFASA when individual differences are minimal.

Theorem 3: CFASA identifiability

Consider the CFASA model with parameter vector $\theta^* \in \Theta$ generating data $\{(\mathbf{x}_i, \eta_i^*)\}_{i=1}^N$. Under Identifiability Conditions (I1)–(I4), the model parameters are locally identifiable up to admissible transformations.

Specifically, there exists a neighborhood $\mathcal{N}(\theta^*) \subset \Theta$ such that if $\theta, \theta' \in \mathcal{N}(\theta^*)$ generate the same probability distribution for X, then:

$$\theta' = T(\theta)$$

where T represents an admissible transformation (rotation, scaling, or permutation that preserves the model structure).

Furthermore, the Fisher Information Matrix $\mathbf{I}(\theta^*)$ is positive definite on the quotient space $\Theta/\sim$ where $\theta_1 \sim \theta_2$ if $T(\theta_1) = \theta_2$ for some admissible transformation T.

Identifiability conditions

(I1) *Factor Structure Identifiability*: The latent factor η^* satisfies standard factor analysis identifiability conditions:

$E[\eta^*] = 0,\ \mathrm{Var}[\eta^*] = 1$ (scale normalization)

The attention-weighted indicators have sufficient rank: $\mathrm{rank}(E[\mathbf{A}^*\mathbf{X}\mathbf{X}^T(\mathbf{A}^*)^T]) = 1$

(I2) *Attention Pattern Diversity*: The population attention patterns exhibit sufficient variation:

$$\mathrm{rank}(\mathrm{Cov}[\mathbf{a}^*(X)]) = p - 1$$

where $\mathbf{a}^*(\mathbf{x}) = \mathrm{softmax}(\mathbf{r}^*(\mathbf{x}))$ are the true attention weights.

(I3) *Non-degeneracy of Encoder*: The encoder function exhibits sufficient nonlinearity:

$$E\left[\left\|\frac{\partial^2 f_{\mathrm{encoder}}(X;\theta^*)}{\partial \theta_{\mathrm{enc}}^2}\right\|_F^2\right] > 0$$

(I4) *Temperature Boundedness*: The temperature parameter is bounded away from extremes:

$$\tau^* \in (\tau_{\min}, \tau_{\max})$$

where $0 < \tau_{\min} < \tau_{\max} < \infty$.

Admissible transformations

The CFASA model admits the following invariance transformations:

T1 (Scale Invariance): For any $c > 0$:

$$(\theta, \eta^*) \mapsto (c^{-1}\theta_{\mathrm{scale}}, c\eta^*)$$

where θ_{scale} represents parameters that affect the factor scale.

T2 (Sign Invariance):

$$(\theta, \eta^*) \mapsto (-\theta_{\mathrm{sign}}, -\eta^*)$$

where θ_{sign} represents parameters that affect the factor sign.

T3 (Permutation Invariance): For permutation matrix $\mathbf{P}$:

$$\mathbf{W}_3 \mapsto \mathbf{P}\mathbf{W}_3, \quad \mathbf{b}_3 \mapsto \mathbf{P}\mathbf{b}_3$$

(only when indicators are exchangeable)

Proof. The proof uses differential geometric methods and the theory of estimating equations.

Step 1: Fisher Information Matrix Structure

The Fisher Information Matrix for CFASA has the block structure:

$$\mathbf{I}(\theta^*) = \begin{pmatrix} \mathbf{I}_{11} & \mathbf{I}_{12} & \mathbf{I}_{13} \\ \mathbf{I}_{12}^T & \mathbf{I}_{22} & \mathbf{I}_{23} \\ \mathbf{I}_{13}^T & \mathbf{I}_{23}^T & \mathbf{I}_{33} \end{pmatrix}$$

where:

$\mathbf{I}_{11}$: Encoder parameters $(\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1, \mathbf{b}_2)$
$\mathbf{I}_{22}$: Loading predictor $(\mathbf{W}_3, \mathbf{b}_3)$
$\mathbf{I}_{33}$: Temperature parameter τ

Step 2: Establish Non-Singularity of Each Block

Block $\boldsymbol{I}_{11}$ (Encoder Parameters): The Fisher information for encoder parameters is:

$$[\mathbf{I}_{11}]_{jk} = E\left[\frac{\partial \log p(\mathbf{x}|\eta^*)}{\partial \theta_j^{\text{enc}}} \frac{\partial \log p(\mathbf{x}|\eta^*)}{\partial \theta_k^{\text{enc}}}\right]$$

Key computation:

$$\frac{\partial \log p(\mathbf{x}|\eta^*)}{\partial \theta_{\text{enc}}} = \frac{\partial}{\partial \theta_{\text{enc}}}\left[\sum_{j=1}^{p} \log\left(\frac{a_j^*(\mathbf{x})\eta^* x_j}{\sigma_j}\right)\right]$$

This requires the chain rule through the attention mechanism:

$$= \sum_{j=1}^{p} \frac{1}{a_j^*} \frac{\partial a_j^*}{\partial \mathbf{c}} \frac{\partial \mathbf{c}}{\partial \theta_{\text{enc}}} \frac{\eta^* x_j}{\sigma_j}$$

Non-singularity: Under condition (I3), the encoder Hessian is non-zero, ensuring $\mathbf{I}_{11} \succ 0$.

Block $\boldsymbol{I}_{22}$ (Loading Parameters):

$$[\mathbf{I}_{22}]_{jk} = E\left[\frac{\partial a_j^*}{\partial \mathbf{W}_3} \frac{\partial a_k^*}{\partial \mathbf{W}_3}\right] E[\eta^{*2}] E[\mathbf{x}\mathbf{x}^T]$$

Non-singularity: Under condition (I2), the attention derivatives have full rank, ensuring $\mathbf{I}_{22} \succ 0$.

Block $\boldsymbol{I}_{33}$ (Temperature):

$$\mathbf{I}_{33} = E\left[\left(\frac{\partial \log p(\mathbf{x}|\eta^*)}{\partial \tau}\right)^2\right]$$

The derivative involves:

$$\frac{\partial a_j^*}{\partial \tau} = -\frac{1}{\tau^2} \frac{\partial}{\partial \mathbf{r}} \text{softmax}(\mathbf{r}^*) \cdot \mathbf{r}^*$$

Non-singularity: Under condition (I4), this is bounded away from zero.

Step 3: Cross-Block Analysis
The cross-terms $\mathbf{I}_{12}, \mathbf{I}_{13}, \mathbf{I}_{23}$ capture parameter interactions. The key insight is that these don't affect identifiability on the quotient space after accounting for admissible transformations.

Encoder-Loading Interaction ($\mathbf{I}_{12}$): Captures how encoder changes affect optimal loading weights.
Encoder-Temperature Interaction ($\mathbf{I}_{13}$): Generally small when temperature is well-separated from extreme values.
Loading-Temperature Interaction ($\mathbf{I}_{23}$): Captures the attention sharpness effect on optimal loadings.

Step 4: Quotient Space Analysis
Define the equivalence relation $\theta_1 \sim \theta_2$ if they generate the same distribution up to admissible transformations.

Tangent Space Decomposition: The parameter space decomposes as:

$$T_{\theta^*}\Theta = \mathcal{M}(\theta^*) \oplus \mathcal{N}(\theta^*)$$

where:

$\mathcal{M}(\theta^*)$: Identifiable directions (orthogonal to invariances)
$\mathcal{N}(\theta^*)$: Non-identifiable directions (tangent to invariances)

Projected Fisher Information: On the quotient space:

$$\tilde{\mathbf{I}}(\theta^*) = \mathbf{P}_{\mathcal{M}}\mathbf{I}(\theta^*)\mathbf{P}_{\mathcal{M}}$$

where $\mathbf{P}_{\mathcal{M}}$ projects onto the identifiable subspace.

Step 5: Verify Positive Definiteness on Quotient Space
Dimension Counting:

Total parameters: $d = \dim(\Theta)$
Invariance constraints: $d_{\text{inv}} = \dim(\mathcal{N}(\theta^*))$
Identifiable parameters: $d_{\text{id}} = d - d_{\text{inv}}$

For CFASA:

Scale invariance: 1 constraint
Sign invariance: 1 constraint
Permutation invariance: 0 constraints (generally not applicable)
Total: $d_{\text{inv}} = 2$

Key Result: Under conditions (I1)–(I4):

$$\text{rank}(\tilde{\mathbf{I}}(\theta^*)) = d_{\text{id}} = d - 2$$

Therefore, $\tilde{\mathbf{I}}(\theta^*) \succ 0$ on the quotient space, establishing local identifiability.

Constructive verification
The identifiability conditions can be verified empirically:

Condition (I1): Check rank of attention-weighted covariance matrix
$\boldsymbol{\Sigma}_{\text{aw}} = E[\mathbf{A}^* \mathbf{X}\mathbf{X}^T (\mathbf{A}^*)^T]$; verify: $\text{rank}(\boldsymbol{\Sigma}_{\text{aw}}) = 1$

Condition (I2): Compute attention pattern covariance
$\text{Cov}_{\mathbf{a}} = \text{Cov}[\mathbf{a}^*(X)]$; verify: $\text{rank}(\text{Cov}_{\mathbf{a}}) = p - 1$

Condition (I3): Check encoder non-degeneracy
$H_{\text{enc}} = E[\| \partial^2 f_{\text{encoder}} / \partial \theta_{\text{enc}}^2 \|_F^2]$; verify: $H_{\text{enc}} > 0$

Condition (I4): Ensure temperature bounds
verify: $\tau^* \in (\tau_{\min}, \tau_{\max}) \subset (0, \infty)$

Implications

1. *Model Specification*: Conditions provide guidance for model architecture choices.

2. *Diagnostic Testing*: Can check identifiability empirically using sample analogs.

3. *Optimization*: Non-identifiable directions explain multiple local minima.

4. *Uncertainty Quantification*: Fisher Information provides standard errors accounting for invariances.

Theorem 4: CFASA parameter consistency

Let $\{(\mathbf{x}_i, \eta_i^*)\}_{i=1}^N$ be i.i.d. observations from the CFASA population model with true parameter vector $\theta^* \in \Theta$. Under Regularity Conditions (C1)–(C5), the M-estimator $\hat{\theta}_N$ satisfies:

$$\hat{\theta}_N \xrightarrow{\text{a.s.}} \theta^* \quad \text{as} N \to \infty$$

and consequently, the attention weight estimators satisfy:

$$\hat{\mathbf{a}}_i(\mathbf{x}_i) \xrightarrow{\text{a.s.}} \mathbf{a}_i^*(\mathbf{x}_i) \quad \text{uniformlyover} i$$

Regularity conditions

(C1) *Compact Parameter Space*: $\Theta \subset \mathbb{R}^d$ is compact,
where $\theta = (\text{vec}(\mathbf{W}_1), \text{vec}(\mathbf{W}_2), \text{vec}(\mathbf{W}_3), \mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3, \tau)$.
(C2) *Continuity*: The loss function $\ell(\theta; \mathbf{x}_i, \eta_i^*)$ is continuous in θ for each $(\mathbf{x}_i, \eta_i^*)$.
(C3) *Uniform Integrability*: $E[\sup_{\theta \in \Theta} |\ell(\theta; X, \eta^*)|] < \infty$.
(C4) *Identification*: The population loss $L(\theta) = E[\ell(\theta; X, \eta^*)]$ has a unique global minimum at θ^*.
(C5) *Attention Distinctiveness*: $\text{Var}[\mathbf{a}(X; \theta^*)] \succ \gamma \mathbf{I}_p$ for some $\gamma > 0$.

Proof. The proof proceeds through the standard M-estimation framework using the uniform law of large numbers.

Step 1: Uniform Convergence of Empirical Loss
Define the empirical loss function:

$$L_N(\theta) = \frac{1}{N} \sum_{i=1}^N \ell(\theta; \mathbf{x}_i, \eta_i^*)$$

We establish uniform convergence: $\sup_{\theta \in \Theta} |L_N(\theta) - L(\theta)| \xrightarrow{\text{a.s.}} 0$.

Verification:
Condition (C1) ensures Θ is totally bounded
Condition (C2) ensures continuity for each observation
Condition (C3) provides the uniform integrability required for the strong law

The empirical process $\{L_N(\theta) - L(\theta) : \theta \in \Theta\}$ forms a Glivenko-Cantelli class by Theorem 2.4.1 in van der Vaart & Wellner (1996), ensuring uniform convergence.

Step 2: Consistency of Argmin
Since $\hat{\theta}_N = \text{argmin}_{\theta \in \Theta} L_N(\theta)$ and $\theta^* = \text{argmin}_{\theta \in \Theta} L(\theta)$, the argmin theorem (Theorem 5.7 in van der Vaart, 1998) gives:

If $L_N \to L$ uniformly and L has a unique minimum, then $\hat{\theta}_N \to \theta^*$.

Verification of Unique Minimum: Condition (C4) directly provides identification.

For CFASA specifically, this requires:

The attention mechanism doesn't collapse to uniform weights (C5)
Response patterns provide information about attention patterns
Factor structure is identifiable through classical rank conditions

Step 3: Attention Weight Convergence
The attention weights are computed as:

$$\hat{\mathbf{a}}_i = \text{softmax}\left(\frac{\mathbf{W}_3 f_{\text{encoder}}(\mathbf{x}_i; \hat{\theta}_N)}{\hat{\tau}_N}\right)$$

Since:

1. f_{encoder} is continuous (ReLU networks are continuous)
2. softmax is continuous on $\mathbb{R}^p$
3. $\hat{\theta}_N \to \theta^*$ almost surely

The continuous mapping theorem yields:

$$\hat{\mathbf{a}}_i(\mathbf{x}_i) \to \mathbf{a}_i^*(\mathbf{x}_i) \quad \text{a.s.}$$

Step 4: Uniform Convergence
For uniform convergence over individuals, we use the fact that the attention function is Lipschitz continuous (from Theorem 1) combined with the compactness of the input space.

Uniform Lipschitz Property: There exists $L > 0$ such that for all $\theta, \theta' \in \Theta$ and all $\mathbf{x}$:

$$\| \mathbf{a}(\mathbf{x}; \theta) - \mathbf{a}(\mathbf{x}; \theta') \|_2 \leq L \| \theta - \theta' \|_2$$

This follows from the Lipschitz continuity of ReLU networks and the softmax function.
Since $\| \hat{\theta}_N - \theta^* \| \to 0$ a.s., we have:

$$\sup_i \| \hat{\mathbf{a}}_i - \mathbf{a}_i^* \|_2 \leq L \| \hat{\theta}_N - \theta^* \|_2 \to 0 \quad \text{a.s.}$$

Verification of regularity conditions for CFASA

(C1): Enforced through bounded network weights: $\| \mathbf{W}_j \|_F \leq M$, $\| \mathbf{b}_j \|_2 \leq B$, $\tau \in [\tau_{\min}, \tau_{\max}]$.
(C2): CFASA loss is continuous as composition of continuous functions (ReLU, softmax, MSE).
(C3): Loss is bounded since softmax outputs are in $(0,1)$ and factors are bounded.
(C4): Follows from attention distinctiveness preventing degeneracy.
(C5): Verified empirically — if all individuals had identical attention patterns, the model would reduce to traditional CFA.

Theorem 5: CFASA asymptotic normality

Under Regularity Conditions (C1)–(C5) from Theorem 4 and additional Smoothness Conditions (S1)–(S4), the CFASA parameter estimator satisfies:

$$\sqrt{N}(\hat{\theta}_N - \theta^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{V})$$

where $\mathbf{V} = \mathbf{H}^{-1}\mathbf{\Omega}\mathbf{H}^{-1}$ with:

$\mathbf{H} = E[\nabla^2 \ell(\theta^*; X, \eta^*)]$ (Hessian matrix)

$\mathbf{\Omega} = \mathrm{Var}[\nabla \ell(\theta^*; X, \eta^*)]$ (Outer product of scores)

Furthermore, for the attention weights:

$$\sqrt{N}(\hat{\mathbf{a}}_i - \mathbf{a}_i^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{G}_i \mathbf{V} \mathbf{G}_i^T)$$

where $\mathbf{G}_i = \nabla_\theta \mathbf{a}_i(\mathbf{x}_i; \theta^*)|_{\theta=\theta^*}$ is the Jacobian matrix.

Additional smoothness conditions

(S1) *Twice Differentiability*: $\ell(\theta; \mathbf{x}, \eta^*)$ is twice continuously differentiable in θ in a neighborhood of θ^*.

(S2) *Non-singular Hessian*: $\mathbf{H} = E[\nabla^2 \ell(\theta^*; X, \eta^*)]$ is positive definite.

(S3) *Finite Moments*: $E[\| \nabla \ell(\theta^*; X, \eta^*) \|^2] < \infty$ and $E[\| \nabla^2 \ell(\theta; X, \eta^*) \|^2] < \infty$ in a neighborhood of θ^*.

(S4) *Attention Jacobian*: $\mathbf{G}_i = \nabla_\theta \mathbf{a}_i(\mathbf{x}_i; \theta)$ exists and is continuous at θ^*.

Proof. The proof follows the classical theory for M-estimators combined with the delta method for nonlinear transformations.

Step 1: Establish M-Estimator Framework

The CFASA estimator solves the first-order condition:

$$\frac{1}{N}\sum_{i=1}^{N} \nabla \ell(\hat{\theta}_N; \mathbf{x}_i, \eta_i^*) = \mathbf{0}$$

Define the score function:

$$\mathbf{S}_N(\theta) = \frac{1}{N}\sum_{i=1}^{N} \nabla \ell(\theta; \mathbf{x}_i, \eta_i^*)$$

By the implicit function theorem and consistency (Theorem 4):

$$\hat{\theta}_N - \theta^* = -\left[\frac{1}{N}\sum_{i=1}^{N} \nabla^2 \ell(\tilde{\theta}_N; \mathbf{x}_i, \eta_i^*)\right]^{-1} \mathbf{S}_N(\theta^*)$$

where $\tilde{\theta}_N$ lies between $\hat{\theta}_N$ and θ^*.

Step 2: Apply Central Limit Theorem

Score Function CLT: Under conditions (S1)–(S3), the score function satisfies:

$$\sqrt{N}\mathbf{S}_N(\theta^*) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \nabla \ell(\theta^*; \mathbf{x}_i, \eta_i^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{\Omega})$$

This follows from the classical CLT since:

$E[\nabla \ell(\theta^*; X, \eta^*)] = \mathbf{0}$ (first-order condition at optimum)

$\mathbf{\Omega} = \mathrm{Var}[\nabla \ell(\theta^*; X, \eta^*)] < \infty$ by (S3)

Hessian Convergence: By the strong law of large numbers and condition (S2):

$$\frac{1}{N}\sum_{i=1}^{N} \nabla^2 \ell(\tilde{\theta}_N; \mathbf{x}_i, \eta_i^*) \xrightarrow{p} \mathbf{H}$$

Step 3: Combine Results for Parameter Asymptotic Normality
From Steps 1–2 and Slutsky's theorem:

$$\sqrt{N}(\hat{\theta}_N - \theta^*) = \mathbf{H}^{-1} \cdot \sqrt{N}\mathbf{S}_N(\theta^*) + o_p(1) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{H}^{-1}\mathbf{\Omega}\mathbf{H}^{-1})$$

Step 4: Delta Method for Attention Weights
The attention weights are nonlinear functions of the parameters:

$$\mathbf{a}_i(\mathbf{x}_i; \theta) = \text{softmax}\left(\frac{\mathbf{W}_3 f_{\text{encoder}}(\mathbf{x}_i;\theta)}{\tau}\right)$$

Jacobian Computation: The Jacobian matrix $\mathbf{G}_i = \nabla_\theta \mathbf{a}_i$ has components:

For encoder parameters $\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1, \mathbf{b}_2$:

$$\frac{\partial \mathbf{a}_i}{\partial \theta_{\text{enc}}} = \frac{1}{\tau}\frac{\partial}{\partial \theta_{\text{enc}}}\left[\text{softmax}(\mathbf{r}_i) \circ \mathbf{W}_3 \frac{\partial \mathbf{c}_i}{\partial \theta_{\text{enc}}}\right]$$

For loading predictor $\mathbf{W}_3, \mathbf{b}_3$:

$$\frac{\partial \mathbf{a}_i}{\partial \mathbf{W}_3} = \frac{1}{\tau}\frac{\partial}{\partial \mathbf{W}_3}\text{softmax}(\mathbf{W}_3\mathbf{c}_i + \mathbf{b}_3)$$

For temperature τ:

$$\frac{\partial \mathbf{a}_i}{\partial \tau} = -\frac{1}{\tau^2}\frac{\partial}{\partial \mathbf{r}}\text{softmax}(\mathbf{r}_i) \cdot \mathbf{r}_i$$

Softmax Derivative: For $\mathbf{s} = \text{softmax}(\mathbf{r})$, the Jacobian is:

$$\frac{\partial s_j}{\partial r_k} = s_j(\delta_{jk} - s_k)$$

where δ_{jk} is the Kronecker delta.
Delta Method Application: Under condition (S4), by the delta method:

$$\sqrt{N}(\hat{\mathbf{a}}_i - \mathbf{a}_i^*) = \mathbf{G}_i\sqrt{N}(\hat{\theta}_N - \theta^*) + o_p(1) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{G}_i\mathbf{V}\mathbf{G}_i^T)$$

Step 5: Explicit Forms for CFASA
CFASA Loss Function:

$$\ell(\theta; \mathbf{x}_i, \eta_i^*) = (\hat{\eta}_i(\theta) - \eta_i^*)^2 + \lambda_{\text{reg}} R(\theta)$$

where $\hat{\eta}_i(\theta) = \mathbf{a}_i(\mathbf{x}_i; \theta)^T \mathbf{x}_i$.

Score Function:

$$\nabla \ell(\theta; \mathbf{x}_i, \eta_i^*) = 2(\hat{\eta}_i - \eta_i^*)\mathbf{G}_i^T\mathbf{x}_i + \lambda_{\text{reg}}\nabla R(\theta)$$

Hessian Matrix:

$$\nabla^2 \ell(\theta; \mathbf{x}_i, \eta_i^*) = 2\mathbf{G}_i^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{G}_i + 2(\hat{\eta}_i - \eta_i^*)\nabla^2\hat{\eta}_i + \lambda_{\text{reg}}\nabla^2 R(\theta)$$

Verification of smoothness conditions

(S1): ReLU networks with finite weights have second derivatives almost everywhere; softmax is C^{∞}.
(S2): Positive definiteness follows from the regularization term and non-degeneracy of attention patterns.
(S3): Finite moments ensured by compact parameter space and bounded loss function.
(S4): Jacobian exists by chain rule; continuity follows from smoothness of component functions.

Implications

1. *Confidence Intervals*: Asymptotic normality enables construction of confidence intervals for attention weights.

2. *Hypothesis Testing*: Can test whether attention patterns differ significantly across individuals or groups.

3. *Model Selection*: Asymptotic theory provides basis for information criteria (AIC, BIC).

4. *Robustness*: The sandwich estimator $\mathbf{H}^{-1}\mathbf{\Omega}\mathbf{H}^{-1}$ is robust to mild model misspecification.

Joint Distribution. For multiple individuals (i, j), the joint asymptotic distribution is:

$$\sqrt{N}\begin{pmatrix}\hat{\mathbf{a}}_i - \mathbf{a}_i^* \\ \hat{\mathbf{a}}_j - \mathbf{a}_j^*\end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \begin{pmatrix}\mathbf{G}_i\mathbf{V}\mathbf{G}_i^T & \mathbf{G}_i\mathbf{V}\mathbf{G}_j^T \\ \mathbf{G}_j\mathbf{V}\mathbf{G}_i^T & \mathbf{G}_j\mathbf{V}\mathbf{G}_j^T\end{pmatrix}\right)$$

This enables testing for differences in attention patterns between individuals.