Arxiv Link

# Method-of-Moments Inference for GLMs

Xingyu Chen, Lin Liu , Rajarshi Mukherjee

Github Link

## Introduction

**Problem Setting and Questions:**

- Samples: $(A_i \in \mathbb{R}, \mathbf{X}_i \in \mathbb{R}^p)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$
- The distribution $\mathbb{P}$ is parameterized as:

  $\mathbf{X} \sim \mathbb{P}_{\mathbf{X}}, \quad \mathsf{E}(A|\mathbf{X}) = \phi(\boldsymbol{\alpha}^\top \mathbf{X}), \quad \mathsf{var}(A|\mathbf{X}) = \sigma^2(\mathbf{X})$

  where $\mathbf{X}$ has mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$
- Asymptotic regime: $\dfrac{p}{n} \to \delta \in [0, +\infty)$ as $n \to \infty$

**Questions**

How can we conduct inference on $\boldsymbol{\alpha}$?
How can we conduct inference on $\|\boldsymbol{\alpha}\|_{\boldsymbol{\Sigma}}^2 := \boldsymbol{\alpha}^\top \boldsymbol{\Sigma} \boldsymbol{\alpha}$?
How can we conduct inference on other functionals?

We focus on four settings, ranging from simple to complex, to introduce our methodology:

- **Case I (Gaussian, known $\boldsymbol{\mu} = 0$ & known $\boldsymbol{\Sigma}$):** $\mathbf{X} \sim \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, $\boldsymbol{\mu} = 0$, $\boldsymbol{\Sigma}$ is known
- **Case II (Gaussian, unknown $\boldsymbol{\mu}$ & known $\boldsymbol{\Sigma}$):** $\mathbf{X} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\mu}$ is unknown, $\boldsymbol{\Sigma}$ is known
- **Case III (Gaussian, unknown $\boldsymbol{\mu}$ & unknown $\boldsymbol{\Sigma}$):** $\mathbf{X} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown
- **Case IV (Missing Data):** $\mathbf{X} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\mu}$ is unknown, $\boldsymbol{\Sigma}$ is known. In addition, we observe $(Y_i)_{i=1}^n$ with $\mathsf{E}[Y_i|\mathbf{X}_i] = \boldsymbol{\beta}^\top \mathbf{X}_i$ and assume $A \perp Y \mid \mathbf{X}$.

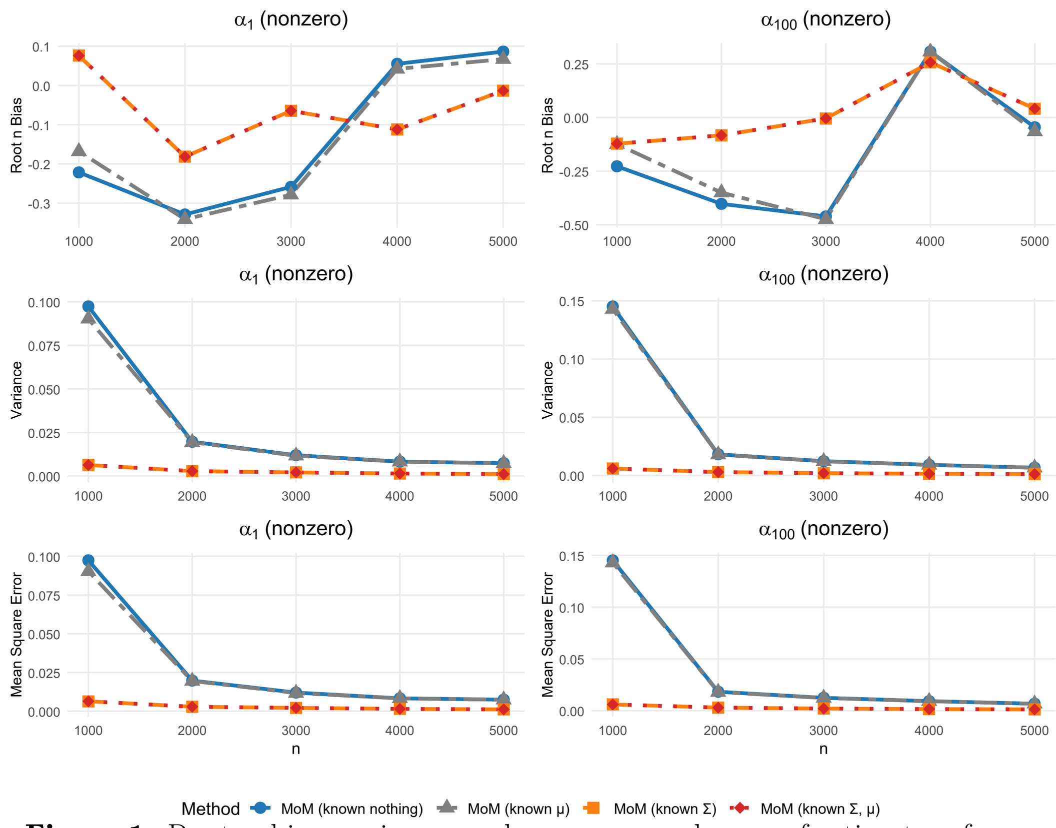


**Figure 1:** Root-$n$ bias, variance, and mean squared error of estimators for $\alpha_1$ and $\alpha_{100}$ under different assumptions on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, with $p/n = 0.4$.

## Estimation in Case I

**Lemma 1 (Stein's Lemma)** *Let $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and let $f : \mathbb{R}^p \to \mathbb{R}$ be a differentiable function such that all expectations below are finite. Then:*

$$\mathbb{E}[\mathbf{X} f(\mathbf{X})] = \boldsymbol{\Sigma}\, \mathbb{E}[\nabla f(\mathbf{X})] + \mathbb{E}[f(\mathbf{X})]\boldsymbol{\mu} \qquad (1)$$

For $f(\mathbf{X}) = \phi(\boldsymbol{\alpha}^\top \mathbf{X})$, we have:

$$\mathbb{E}[\mathbf{X}A] = \mathbb{E}[\mathbf{X}\,\phi(\boldsymbol{\alpha}^\top \mathbf{X})] = \mathbb{E}[\phi'(\mathbf{Z})]\boldsymbol{\Sigma}\,\boldsymbol{\alpha} + \mathbb{E}[f(\mathbf{X})]\boldsymbol{\mu} \qquad (2)$$

Here and below, $\mathbf{Z} \sim \boldsymbol{\alpha}^\top \mathbf{X} \sim \mathcal{N}(\boldsymbol{\alpha}^\top \boldsymbol{\mu}, \|\boldsymbol{\alpha}\|_{\boldsymbol{\Sigma}}^2)$, with $\lambda_\alpha := \boldsymbol{\alpha}^\top \boldsymbol{\mu}$ and $\gamma_\alpha^2 := \|\boldsymbol{\alpha}\|_{\boldsymbol{\Sigma}}^2$. We define $\mathrm{f}_i(\lambda_\alpha, \gamma_\alpha^2) := \mathbb{E}[\phi^{(i)}(\mathbf{Z})]$.
Since $\boldsymbol{\mu} = \mathbf{0}$ in Case I:

**Identification Equations for $\boldsymbol{\mu} = \mathbf{0}$**

$$\begin{aligned} \mathrm{m}_{\alpha_j} &:= \mathbb{E}[A\mathbf{X}^\top]\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j = \mathrm{f}_1(0, \gamma_\alpha^2)\boldsymbol{\alpha}_j, \\ \mathrm{m}_{\mathbf{X}A,2} &:= \mathbb{E}[A_1 \mathbf{X}_1^\top \boldsymbol{\Sigma}^{-1} \mathbf{X}_2 A_2] = \mathrm{f}_1(0, \gamma_\alpha^2)^2 \cdot \gamma_\alpha^2 =: \Psi(\gamma_\alpha^2) \end{aligned} \qquad (3)$$

Since $\boldsymbol{\Sigma}$ is **known** in Case I, we construct $U$-statistics to unbiasedly estimate the required moments:

**Moment Estimators for $\boldsymbol{\mu} = 0$**

$$\begin{aligned} \widehat{\mathrm{m}_{\alpha_j}} &:= \frac{1}{n}\sum_{1 \le i \le n} A_i \mathbf{X}_i^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j \\ &=: \mathbb{U}_{n,1}[A_1 \mathbf{X}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j], \\ \widehat{\mathrm{m}_{\mathbf{X}A,2}} &:= \frac{1}{n(n-1)} \sum_{1 \le i_1 \ne i_2 \le n} A_{i_1} \mathbf{X}_{i_1}^\top \boldsymbol{\Sigma}^{-1} \mathbf{X}_{i_2} A_{i_2} \\ &=: \mathbb{U}_{n,2}[A_1 \mathbf{X}_1^\top \boldsymbol{\Sigma}^{-1} \mathbf{X}_2 A_2] \end{aligned} \qquad (4)$$

We then plug the moment estimators from (4) into the identification equations (3) and solve the system to obtain the estimator for the parameter of interest.

$$\begin{aligned} \widehat{\gamma_\alpha^2} &= \Psi^{-1}(\widehat{\mathrm{m}_{\mathbf{X}A,2}}) \\ \widehat{\boldsymbol{\alpha}_j} &= \frac{\widehat{\mathrm{m}_{\alpha_j}}}{\mathrm{f}_1(0, \widehat{\gamma_\alpha^2})} \end{aligned} \qquad (5)$$

## Estimation in Case II

Since $\boldsymbol{\mu}$ is unknown in Case II, the identification equations are as follows:

**Identification Equations for Unknown $\boldsymbol{\mu}$**

$$\begin{aligned} m_A &:= \mathbb{E}[A] = \mathrm{f}_0(\lambda_\alpha, \gamma_\alpha^2), \\ m_{\mathbf{X},2} &:= \mathbb{E}[\mathbf{X}^\top]\boldsymbol{\Sigma}^{-1}\mathbb{E}[\mathbf{X}] = \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \\ m_{\mathbf{X}A,\mathbf{X}} &:= \mathbb{E}[A\mathbf{X}^\top]\boldsymbol{\Sigma}^{-1}\mathbb{E}[\mathbf{X}] \\ &= m_A \cdot m_{\mathbf{X},2} + \mathrm{f}_1(\lambda_\alpha, \gamma_\alpha^2) \cdot \lambda_\alpha, \\ m_{\mathbf{X}A,2} &:= \mathbb{E}[A\mathbf{X}^\top]\boldsymbol{\Sigma}^{-1}\mathbb{E}[A\mathbf{X}] \\ &= m_A^2 \cdot m_{\mathbf{X},2} + \mathrm{f}_1^2(\lambda_\alpha, \gamma_\alpha^2) \cdot \gamma_\alpha^2 + \\ &\quad 2 \cdot m_A \cdot \mathrm{f}_1(\lambda_\alpha, \gamma_\alpha^2) \cdot \lambda_\alpha, \\ m_{\nu_j} &:= \mathbb{E}[\mathbf{X}]^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j = \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j = \boldsymbol{\nu}^\top \boldsymbol{e}_j = \nu_j, \\ m_{\alpha_j} &:= \mathbb{E}[A\mathbf{X}^\top]\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j \\ &= \mathrm{f}_0(\lambda_\alpha, \gamma_\alpha^2) \cdot \nu_j + \mathrm{f}_1(\lambda_\alpha, \gamma_\alpha^2) \cdot \boldsymbol{\alpha}_j. \end{aligned} \qquad (6)$$

The first four moments in (6) can be reduced to two equations, which form a diffeomorphism map $\Psi_{GLM}$:

**Reduced Identification Equations for Unknown $\boldsymbol{\mu}$**

$$\begin{aligned} m_1 &:= m_A = \mathrm{f}_0(\lambda_\alpha, \gamma_\alpha^2), \\ m_2 &:= m_{\mathbf{X}A,2} + m_A^2 \cdot m_{\mathbf{X},2} - 2 \cdot m_A \cdot m_{\mathbf{X}A,\mathbf{X}} \\ &= \mathrm{f}_1^2(\lambda_\alpha, \gamma_\alpha^2) \cdot \gamma_\alpha^2, \\ \Psi_{GLM} &: (\lambda_\alpha, \gamma_\alpha^2) \to (m_1, m_2). \end{aligned} \qquad (7)$$

Since $\boldsymbol{\Sigma}$ is **known** in Case II:

**Moment Estimators for Unknown $\boldsymbol{\mu}$**

$$\begin{aligned} \widehat{m}_1 &:= \widehat{m}_A := \mathbb{U}_{n,1}[A], \\ \widehat{m}_2 &:= \widehat{m}_{\mathbf{X}A,2} + \widehat{m}_A^2 \cdot \widehat{m}_{\mathbf{X},2} - 2 \cdot \widehat{m}_A \cdot \widehat{m}_{\mathbf{X}A,\mathbf{X}}, \\ &\text{where } \ \widehat{m}_{\mathbf{X},2} := \mathbb{U}_{n,2}[\mathbf{X}_1^\top \boldsymbol{\Sigma}^{-1}\mathbf{X}_2], \\ \widehat{m}_{\mathbf{X}A,\mathbf{X}} &:= \mathbb{U}_{n,2}[A_1 \mathbf{X}_1^\top \boldsymbol{\Sigma}^{-1}\mathbf{X}_2], \\ \widehat{m}_{\mathbf{X}A,2} &:= \mathbb{U}_{n,2}[A_1 \mathbf{X}_1^\top \boldsymbol{\Sigma}^{-1}\mathbf{X}_2 A_2], \\ \widehat{m}_{\nu_j} &:= \mathbb{U}_{n,1}[\mathbf{X}^\top]\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j, \\ \widehat{m}_{\alpha_j} &:= \mathbb{U}_{n,1}[A\mathbf{X}^\top]\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j. \end{aligned} \qquad (8)$$

We then plug the moment estimators from (8) into the reduced identification equations (7):

$$\begin{aligned} (\widehat{\lambda_\alpha}, \widehat{\gamma_\alpha^2}) &:= \Psi_{GLM}^{-1}(\widehat{m_1}, \widehat{m_2}), \\ \widehat{\boldsymbol{\alpha}}_j &:= \frac{\widehat{m}_{\alpha_j} - \mathrm{f}_0(\widehat{\lambda}_\alpha, \widehat{\gamma_\alpha^2}) \cdot \widehat{m}_{\nu_j}}{\mathrm{f}_1(\widehat{\lambda}_\alpha, \widehat{\gamma_\alpha^2})}. \end{aligned} \qquad (9)$$

## Estimation in Case III

In Case III, the identification equations remain the same as in Case II, i.e.,

**Identification Equations for Unknown $\boldsymbol{\mu}$**

Same as (7).

The knowledge of $\boldsymbol{\Sigma}$ influences the construction of moment estimators. We propose two methods to address this, each with theoretical guarantees.
One method involves using a sample splitting strategy with weighted sample covariance:

**Moment Estimators Unknown $\boldsymbol{\Sigma}$ (Sample Splitting)**

$$\begin{aligned} \widetilde{\boldsymbol{\Sigma}} &:= \frac{1}{\frac{n}{2} - p - 1}\sum_{j \in I_2}(\mathbf{X}_j - \bar{\mathbf{X}}_{I_2})(\mathbf{X}_j - \bar{\mathbf{X}}_{I_2})^\top, \\ &\text{where } \bar{\mathbf{X}}_{I_2} := \frac{1}{n/2}\sum_{j \in I_2}\mathbf{X}_j, \\ \widehat{m}_{\mathbf{X}A,2} &:= \frac{1}{\frac{n}{2}\left(\frac{n}{2} - 1\right)} \sum_{i_1 \ne i_2 \in I_1} A_{i_1}\mathbf{X}_{i_1}^\top \widetilde{\boldsymbol{\Sigma}}^{-1}\mathbf{X}_{i_2}A_{i_2}. \end{aligned} \qquad (10)$$

Another method uses Chebyshev polynomials to approximate $\boldsymbol{\Sigma}^{-1}$:

**Moment Estimators Unknown $\boldsymbol{\Sigma}$ (Chebyshev)**

$$\begin{aligned} \boldsymbol{\Sigma}^{-1} &\approx \sum_{l=0}^{J} c_l \boldsymbol{\Sigma}^l, \\ \widehat{m}_{\mathbf{X}A,2} &:= \sum_{l=0}^{J} c_l \mathbb{U}_{n,l+2}\left[A_1\mathbf{X}_1^\top\left(\prod_{s=3}^{l+2}\mathbf{X}_s\mathbf{X}_s^\top\right)\mathbf{X}_2 A_2\right]. \end{aligned} \qquad (11)$$

## Estimation in Case IV

Now, we apply the methods discussed above to the Missing Data setting, also considered in [1].
Here, the observed data are $(A_i, X_i, A_i \cdot Y_i)_{i=1}^n$, and the parameter of interest is $\psi := \mathbb{E}[Y] = \mathbb{E}[\boldsymbol{\beta}^\top \mathbf{X}] = \boldsymbol{\beta}^\top \boldsymbol{\mu}$.

**Identification Equations**

Equations in (7), plus two additional equations:

$$\begin{aligned} m_{AY} &:= \mathbb{E}[AY] = m_A \cdot \psi + \mathrm{f}_1(\lambda_\alpha, \gamma_\alpha^2) \cdot \gamma_{\alpha,\beta}, \\ m_{\mathbf{X}AY,\mathbf{X}} &:= \mathbb{E}[YA\mathbf{X}^\top]\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ &= (m_A + m_{\mathbf{X}A,\mathbf{X}}) \cdot \psi + \\ &\quad \{m_{\mathbf{X},2} \cdot \mathrm{f}_1(\lambda_\alpha, \gamma_\alpha^2) + \mathrm{f}_2(\lambda_\alpha, \gamma_\alpha^2) \cdot \lambda_\alpha\} \cdot \gamma_{\alpha,\beta}, \\ &\text{where } \ \gamma_{\alpha,\beta} := \boldsymbol{\alpha}^\top \boldsymbol{\Sigma} \boldsymbol{\beta}. \end{aligned} \qquad (12)$$

As in [1], we use the knowledge of $\boldsymbol{\Sigma}$, so the additional moment estimators are the same as those in (8), and the final estimator follows similarly. Comparison below:

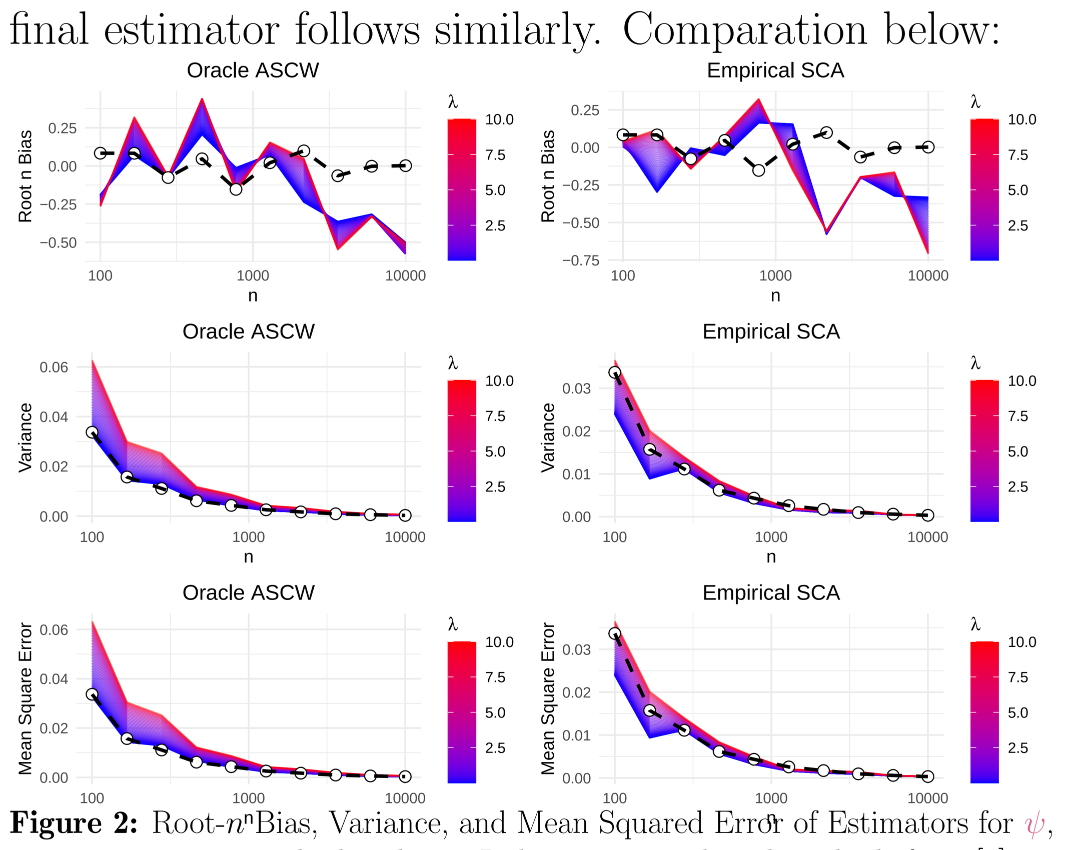


**Figure 2:** Root-$n$ Bias, Variance, and Mean Squared Error of Estimators for $\psi$, comparing our method with two Ridge regression-based methods from [1].

**Theorem (Informal), Gaussian, CAN**

When $\boldsymbol{\Sigma}$ is known, under some mild conditions, the above estimators are all $\sqrt{n}$-consistent.

Further assume that $\sqrt{p}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}$, $\sqrt{p}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\alpha}$, and $\sqrt{p}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\beta}$ and their inner products with respect to $\boldsymbol{\Sigma}^{-1}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}^2$, $\boldsymbol{\Sigma}^3$ converge to a nontrivial distribution. Then, the above estimators, after scaling by $\sqrt{n}$, converge to a normal distribution.

When $\boldsymbol{\Sigma}$ is unknown, for the method in (10), which requires $\frac{n}{2} > p + 3$, we can show that our method is $\sqrt{n}$-consistent. For the method in (11), we can show that our method is consistent.

**Theorem (Informal), Universality**

When $\mathbf{X}$ violates the Gaussian distribution, but $\boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$ has zero mean and unit variance, the above identification equations (3), (6) hold with error $\mathcal{O}\left(n^{-3/4}\right)$.

Thus, the above consistent results for the Gaussian model will still hold.

**Variance Estimator**

We use a bootstrap method to estimate the variance of the U-statistics and then apply the Delta method to estimate the variance of the parameters of interest.

## References

[1] Michael Celentano and Martin J Wainwright. Challenges of the inconsistency regime: Novel debiasing methods for missing data models. *arXiv preprint arXiv:2309.01362*, 2023.

poster