Supplemental Information: Graphical Models for Non-Negative Data Using Generalized Score Matching

Shiqing Yu
University of Washington

Mathias Drton
University of Washington

Ali Shojaie
University of Washington

A.2 SCORE MATCHING

The following lemma is used in the proof of Theorem 2.

Lemma A.1. *Assuming that f and g are differentiable a.e., then for all $j = 1, \ldots, m$,*

$$\lim_{a\nearrow+\infty, b\searrow 0^+} f(\boldsymbol{x}_{-j};a)g(\boldsymbol{x}_{-j};a) - f(\boldsymbol{x}_{-j};b)g(\boldsymbol{x}_{-j};b) = \int_0^\infty f(\boldsymbol{x})\frac{\partial g(\boldsymbol{x})}{\partial x_j}\,\mathrm{d}x_j + \int_0^\infty g(\boldsymbol{x})\frac{\partial f(\boldsymbol{x})}{\partial x_j}\,\mathrm{d}x_j,$$

where $(\boldsymbol{x}_{-j};a)$ is the vector obtained by replacing the j-th component of $\boldsymbol{x}$ by a.

Proof. This is just an analog of Lemma 4 from Hyvärinen (2005) proved by integrating the partial derivatives. □

Proof of Theorem 2. Recall the following assumptions given in Section 2.3.

$$\begin{aligned}
&\text{(A1)}\quad p_0(\boldsymbol{x})h_j(x_j)\partial_j \log p(\boldsymbol{x}) \to 0 \text{ as } x_j \nearrow +\infty \text{ and as } x_j \searrow 0^+,\ \forall \boldsymbol{x}_{-j} \in \mathbb{R}_+^{m-1},\ \forall p \in \mathcal{P}_+,\\
&\text{(A2)}\quad \mathbb{E}_{p_0}\|\nabla \log p(\boldsymbol{X}) \circ \boldsymbol{h}^{1/2}(\boldsymbol{X})\|_2^2 < +\infty, \quad \mathbb{E}_{p_0}\|(\nabla \log p(\boldsymbol{X}) \circ \boldsymbol{h}(\boldsymbol{X}))'\|_1 < +\infty, \quad \forall p \in \mathcal{P}_+,
\end{aligned}$$

where

$$\partial_j \log p(\boldsymbol{x}) \equiv \left.\frac{\partial \log p(\boldsymbol{y})}{\partial y_j}\right|_{\boldsymbol{y}=\boldsymbol{x}}.$$

Without explicitly writing the domains $\mathbb{R}_+$ or $\mathbb{R}_+^m$ in all integrals, by (4) we have

$$\begin{aligned}
J_{\boldsymbol{h}}(p) &= \frac{1}{2}\int p_0(\boldsymbol{x})\Big[\|\nabla \log p(\boldsymbol{x}) \circ \boldsymbol{h}^{1/2}(\boldsymbol{x})\|_2^2 - 2(\nabla \log p(\boldsymbol{x}) \circ \boldsymbol{h}^{1/2}(\boldsymbol{x}))^\top(\nabla \log p_0(\boldsymbol{x}) \circ \boldsymbol{h}^{1/2}(\boldsymbol{x}))\\
&\qquad + \|\nabla \log p_0(\boldsymbol{x}) \circ \boldsymbol{h}^{1/2}(\boldsymbol{x})\|_2^2\Big]\,\mathrm{d}\boldsymbol{x}\\
&= \underbrace{\frac{1}{2}\int p_0(\boldsymbol{x})\sum_{j=1}^m h_j(x_j)\left(\frac{\partial \log p(\boldsymbol{x})}{\partial x_j}\right)^2 \mathrm{d}\boldsymbol{x}}_{\equiv A} \underbrace{- \int p_0(\boldsymbol{x})\sum_{j=1}^m h_j(x_j)\frac{\partial \log p(\boldsymbol{x})}{\partial x_j}\frac{\partial \log p_0(\boldsymbol{x})}{\partial x_j}\,\mathrm{d}\boldsymbol{x}}_{\equiv B}\\
&\qquad + \underbrace{\frac{1}{2}\int p_0(\boldsymbol{x})\sum_{j=1}^m h_j(x_j)\left(\frac{\partial \log p_0(\boldsymbol{x})}{\partial x_j}\right)^2 \mathrm{d}\boldsymbol{x}}_{\equiv C},
\end{aligned}$$

where A will simply appear in the final display as is, C is a constant as it only involves the true pdf p_0, and we wish to simplify B by integration by parts. We can split the integral into these three parts since A and C are assumed finite in the first part of (A2), and the integrand in B is integrable since $|2ab| \le a^2 + b^2$. Thus, by linearity and Fubini's theorem, we can write

$$
\begin{aligned}
B &= -\sum_{j=1}^{m} \int p_0(\boldsymbol{x}) h_j(x_j) \frac{\partial \log p(\boldsymbol{x})}{\partial x_j} \frac{\partial \log p_0(\boldsymbol{x})}{\partial x_j} \,\mathrm{d}\boldsymbol{x} \\
&= -\sum_{j=1}^{m} \int \left[\int p_0(\boldsymbol{x}) h_j(x_j) \frac{\partial \log p(\boldsymbol{x})}{\partial x_j} \frac{\partial \log p_0(\boldsymbol{x})}{\partial x_j} \,\mathrm{d}x_j \right] \mathrm{d}\boldsymbol{x}_{-j}.
\end{aligned}
$$

By the fact that $\frac{\partial \log p_0(\boldsymbol{x})}{\partial x_j} = \frac{1}{p_0(\boldsymbol{x})} \frac{\partial p_0(\boldsymbol{x})}{\partial x_j}$, this can be simplified to

$$
B = -\sum_{j=1}^{m} \int \left[\int \frac{\partial p_0(\boldsymbol{x})}{\partial x_j} h_j(x_j) \frac{\partial \log p(\boldsymbol{x})}{\partial x_j} \,\mathrm{d}x_j \right] \mathrm{d}\boldsymbol{x}_{-j}.
$$

Then by Lemma A.1 and assumption (A1),

$$
\begin{aligned}
B = -\sum_{j=1}^{m} \int \Bigg[& \lim_{a \nearrow \infty, b \searrow 0^+} \left[p_0(\boldsymbol{x}_{-j}; a) h_j(a) \partial_j \log p(\boldsymbol{x}_{-j}, a) - p_0(\boldsymbol{x}_{-j}; b) h_j(b) \partial_j \log p(\boldsymbol{x}_{-j}, b) \right] \\
& - \int p_0(\boldsymbol{x}) \frac{\partial \left(h_j(x_j) \partial_j \log p(\boldsymbol{x}) \right)}{\partial x_j} \,\mathrm{d}x_j \Bigg] \mathrm{d}\boldsymbol{x}_{-j} \\
= \sum_{j=1}^{m} \int \Bigg[& \int p_0(\boldsymbol{x}) \frac{\partial (h_j(x_j) \partial_j \log p(\boldsymbol{x}))}{\partial x_j} \,\mathrm{d}x_j \Bigg] \mathrm{d}\boldsymbol{x}_{-j}.
\end{aligned}
$$

Justified by the second half of (A2), by Fubini-Tonelli and linearity again

$$
\begin{aligned}
B &= \sum_{j=1}^{m} \int p_0(\boldsymbol{x}) \frac{\partial (h_j(x_j) \partial_j \log p(\boldsymbol{x}))}{\partial x_j} \,\mathrm{d}\boldsymbol{x}, \\
&= \sum_{j=1}^{m} \int h_j'(x_j) \frac{\partial \log p(\boldsymbol{x})}{\partial x_j} p_0(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} + \sum_{j=1}^{m} \int h_j(x_j) \frac{\partial^2 \log p(\boldsymbol{x})}{\partial x_j^2} p_0(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
J_{\boldsymbol{h}}(p) &= B + A + C \\
&= \int_{\mathbb{R}_+^m} p_0(\boldsymbol{x}) \sum_{j=1}^{m} \left[h_j'(x_j) \frac{\partial \log p(\boldsymbol{x})}{\partial x_j} + h_j(x_j) \frac{\partial^2 \log p(\boldsymbol{x})}{\partial x_j^2} + \frac{1}{2} h_j(x_j) \left(\frac{\partial \log p(\boldsymbol{x})}{\partial x_j} \right)^2 \right] \mathrm{d}\boldsymbol{x} + C,
\end{aligned}
$$

where C is a constant that does not depend on p. □

Proof of Theorem 3. By definition $J_{\boldsymbol{h}}(p_{\boldsymbol{\theta}}) \geq 0$ and $J_{\boldsymbol{h}}(p_{\boldsymbol{\theta}_0}) = 0$, so $\boldsymbol{\theta}_0$ minimizes $J_{\boldsymbol{h}}(p_{\boldsymbol{\theta}})$. Conversely, suppose $J_{\boldsymbol{h}}(p_{\boldsymbol{\theta}}) = 0$ for some $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}$. By assumption $p_0(\boldsymbol{x}) > 0$ almost surely (hereafter a.s.) and $h_j^{1/2}(\boldsymbol{x}) > 0$ a.s. for all $j = 1, \ldots, m$. Therefore, we must have $\nabla \log p_{\boldsymbol{\theta}_1}(\boldsymbol{x}) = \nabla \log p_0(\boldsymbol{x})$ a.s., or equivalently, $p_{\boldsymbol{\theta}_1}(\boldsymbol{x}) = \text{const} \times p_0(\boldsymbol{x})$ for all almost every $\boldsymbol{x} \in \mathbb{R}_+^m$. Since $p_{\boldsymbol{\theta}_1}$ and p_0 are both continuous probability density functions, we necessarily have $p_{\boldsymbol{\theta}_1}(\boldsymbol{x}) = p_0(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}_+^m$, which implies $\boldsymbol{\theta}_{\boldsymbol{1}} = \boldsymbol{\theta}_0$ by the identifiability assumption. The last claim follows by the law of large numbers, and is an analog of Corollary 3 in Hyvärinen (2005). □

A.3 EXPONENTIAL FAMILIES

Consider the case where $\{p_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^r\}$ contains exponential families with densities

$$
\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^\top \boldsymbol{t}(\boldsymbol{x}) - \psi(\boldsymbol{\theta}) + b(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{R}_+^m.
$$

Then the empirical generalized $\boldsymbol{h}$-score matching loss becomes

$$
\hat{J}_{\boldsymbol{h}}(p_{\boldsymbol{\theta}}) = \frac{1}{2} \boldsymbol{\theta}^\top \boldsymbol{\Gamma}(\mathrm{x}) \boldsymbol{\theta} - \boldsymbol{g}(\mathrm{x})^\top \boldsymbol{\theta} + \text{const},
$$

where

$$\mathbf{\Gamma}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} h_j(X_j^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}^{(i)})^\top \in \mathbb{R}^{r\times r} \quad \text{and} \tag{A.1}$$

$$\boldsymbol{g}(\mathbf{x}) = -\frac{1}{n}\sum_{i=1}^{n}\left[h_j(X_j^{(i)})b_j'(\boldsymbol{X}^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}^{(i)}) + h_j(X_j^{(i)})\boldsymbol{t}_j''(\boldsymbol{X}^{(i)}) + h_j'(X_j^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}_i)\right] \in \mathbb{R}^r. \tag{A.2}$$

Proof of (6). For exponential families, under the assumptions the empirical loss $\hat{J}_{\boldsymbol{h}}(p_{\boldsymbol{\theta}})$ becomes (up to an additive constant)

$$\begin{aligned}
&\hat{J}_{\boldsymbol{h}}(p_{\boldsymbol{\theta}}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\left[h_j'(X_j^{(i)})\frac{\partial \log p_{\boldsymbol{\theta}}(\boldsymbol{X}^{(i)})}{\partial X_j^{(i)}} + h_j(X_j^{(i)})\frac{\partial^2 \log p_{\boldsymbol{\theta}}(\boldsymbol{X}^{(i)})}{\partial (X_j^{(i)})^2} + \frac{1}{2}h_j(X_j^{(i)})\left(\frac{\partial \log p_{\boldsymbol{\theta}}(\boldsymbol{X}^{(i)})}{\partial X_j^{(i)}}\right)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\Big[h_j'(X_j^{(i)})(\boldsymbol{\theta}^\top \boldsymbol{t}_j'(\boldsymbol{X}^{(i)}) + b_j'(\boldsymbol{X}^{(i)})) + h_j(X_j^{(i)})(\boldsymbol{\theta}^\top \boldsymbol{t}_j''(\boldsymbol{X}^{(i)}) + b_j''(\boldsymbol{X}^{(i)})) \\
&\qquad\qquad + \frac{1}{2}h_j(X_j^{(i)})(\boldsymbol{\theta}^\top \boldsymbol{t}_j'(\boldsymbol{X}^{(i)}) + b_j'(\boldsymbol{X}^{(i)}))^2\Big] \\
&= \frac{1}{n}\left\{\frac{1}{2}\boldsymbol{\theta}^\top\left[\sum_{i=1}^{n}\sum_{j=1}^{m} h_j(X_j^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}^{(i)})^\top\right]\boldsymbol{\theta}\right. \\
&\qquad \left. + \left[\sum_{i=1}^{n} h_j(X_j^{(i)})b_j'(\boldsymbol{X}^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}^{(i)}) + h_j(X_j^{(i)})\boldsymbol{t}_j''(\boldsymbol{X}^{(i)}) + h_j'(X_j^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}^{(i)})\right]^\top \boldsymbol{\theta}\right\} + \text{const},
\end{aligned}$$

which is quadratic in $\boldsymbol{\theta}$. Let

$$\mathbf{\Gamma}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} h_j(X_j^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}^{(i)})^\top, \tag{A.3}$$

$$\boldsymbol{g}(\mathbf{x}) = -\frac{1}{n}\sum_{i=1}^{n}\left[h_j(X_j^{(i)})b_j'(\boldsymbol{X}^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}^{(i)}) + h_j(X_j^{(i)})\boldsymbol{t}_j''(\boldsymbol{X}^{(i)}) + h_j'(X_j^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}_i)\right]. \tag{A.4}$$

Then we can write $\hat{J}_{\boldsymbol{h}}(p_{\boldsymbol{\theta}}) = \frac{1}{2}\boldsymbol{\theta}^\top\mathbf{\Gamma}(\mathbf{x})\boldsymbol{\theta} - \boldsymbol{g}(\mathbf{x})^\top\boldsymbol{\theta} + \text{const}$. □

Proof of Theorem 4. Recall that $\hat{J}_{\boldsymbol{h}}(p_{\boldsymbol{\theta}}) = \frac{1}{2}\boldsymbol{\theta}^\top\mathbf{\Gamma}\boldsymbol{\theta} - \boldsymbol{g}^\top\boldsymbol{\theta} + \text{const}$. The minimizer of $\hat{J}_{\boldsymbol{h}}(p_{\boldsymbol{\theta}})$ is thus available in the unique closed form $\hat{\boldsymbol{\theta}} \equiv \mathbf{\Gamma}(\mathbf{x})^{-1}\boldsymbol{g}(\mathbf{x})$ as long as $\mathbf{\Gamma}$ is invertible (C1). Since $\mathbf{\Gamma}$ and $\boldsymbol{g}$ are sample averages, by Khinchin's weak law of large numbers we have $\mathbf{\Gamma} \to_p \mathbb{E}_{p_0}\mathbf{\Gamma} \equiv \mathbf{\Gamma}_0$ and $\boldsymbol{g} \to_p \mathbb{E}_{p_0}\boldsymbol{g} \equiv \boldsymbol{g}_0$, where existence of $\mathbf{\Gamma}_0$ and $\boldsymbol{g}_0$ is assumed in (C2). Since $J_{\boldsymbol{h}}(p_{\boldsymbol{\theta}}) = \mathbb{E}[\hat{J}_{\boldsymbol{h}}(p_{\boldsymbol{\theta}})] = \mathbb{E}[\frac{1}{2}\boldsymbol{\theta}^\top\mathbf{\Gamma}(\mathbf{x})\boldsymbol{\theta} - \boldsymbol{g}(\mathbf{x})^\top\boldsymbol{\theta}] = \frac{1}{2}\boldsymbol{\theta}^\top\mathbf{\Gamma}_0\boldsymbol{\theta} - \boldsymbol{g}_0\boldsymbol{\theta}$ and we know $\boldsymbol{\theta}_0$ minimizes $J_{\boldsymbol{h}}(p_{\boldsymbol{\theta}})$ by definition, by first-order condition we munst have $\mathbf{\Gamma}_0\boldsymbol{\theta}_0 = \boldsymbol{g}_0$. Then by Lindeberg-Lévy central limit theorem (recall that $\boldsymbol{g}(\mathbf{x})$ and $\mathbf{\Gamma}(\mathbf{x})$ are sample averages)

$$\sqrt{n}(\boldsymbol{g}(\mathbf{x}) - \mathbf{\Gamma}(\mathbf{x})\boldsymbol{\theta}_0) \to_d \mathcal{N}_m(\mathbf{0}, \mathbf{\Sigma}_0),$$

where $\mathbf{\Sigma}_0 \equiv \mathbb{E}_{p_0}[(\mathbf{\Gamma}(\mathbf{x})\boldsymbol{\theta}_0 - \boldsymbol{g}(\mathbf{x}))(\mathbf{\Gamma}(\mathbf{x})\boldsymbol{\theta}_0 - \boldsymbol{g}(\mathbf{x}))^\top]$, as long as $\mathbf{\Sigma}_0$ exists (C2).
Then by Slutsky's theorem,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \equiv \sqrt{n}(\mathbf{\Gamma}(\mathbf{x})^{-1}(\boldsymbol{g}(\mathbf{x}) - \mathbf{\Gamma}(\mathbf{x})\boldsymbol{\theta}_0)) \to_d \mathcal{N}_r(\mathbf{0}, \mathbf{\Gamma}_0^{-1}\mathbf{\Sigma}\mathbf{\Gamma}_0^{-1}),$$

as long as $\mathbf{\Gamma}_0$ is invertible (C2).
For the second half of the theorem, (C2) $\mathbb{E}_{p_0}\mathbf{\Gamma}(\mathbf{x}) < \infty$ and $\mathbb{E}_{p_0}\boldsymbol{g}(\mathbf{x}) < \infty$ implies $\mathbb{E}_{p_0}|\mathbf{\Gamma}(\mathbf{x})| < \infty$ and $\mathbb{E}_{p_0}|\boldsymbol{g}(\mathbf{x})| < \infty$, so by strong law of large numbers (and a union bound on at most k^2 null sets)

$$\mathbf{\Gamma}(\mathbf{x}) \to_{\text{a.s.}} \mathbf{\Gamma}_0, \quad \boldsymbol{g}(\mathbf{x}) \to_{\text{a.s.}} \boldsymbol{g}_0.$$

Then outside a null set,

$$\hat{\boldsymbol{\theta}} \equiv \mathbf{\Gamma}(\mathbf{x})^{-1}\boldsymbol{g}(\mathbf{x}) \rightarrow_{\text{a.s.}} \mathbf{\Gamma}_0^{-1}\boldsymbol{g}_0 = \boldsymbol{\theta}_0.$$

$\square$

Proof for Example 5. For the family of univariate truncated Gaussian distributions with unknown mean parameter μ and known variance parameter σ^2, we have

$$p_\theta(x) \propto \exp\left(\theta t(x) + b(x)\right), \quad \theta \equiv \frac{\mu}{\sigma^2}, \quad t(x) \equiv x, \quad b(x) = -\frac{x^2}{2\sigma^2}.$$

We choose to estimate $\theta \equiv \mu/\sigma^2$. Then by (A.1) and (A.2),

$$\begin{aligned}
\hat{\mu}_h &= \sigma^2\hat{\theta} \equiv \sigma^2\Gamma(\boldsymbol{x})^{-1}g(\boldsymbol{x}) \\
&= -\sigma^2\left[\sum_{i=1}^n h(X_i)t'(X_i)^2\right]^{-1}\left[\sum_{i=1}^n h(X_i)b'(X_i)t'(X_i) + h(X_i)t''(X_i) + h'(X_i)t'(X_i)\right] \\
&= -\sigma^2\left[\sum_{i=1}^n h(X_i)\right]^{-1}\left[\sum_{i=1}^n -h(X_i)\frac{X_i}{\sigma^2} + h'(X_i)\right].
\end{aligned}$$

By Theorem 4,

$$\sqrt{n}(\hat{\mu}_h - \mu_0) \rightarrow_d \mathcal{N}\left(0, \frac{\sigma^4\mathbb{E}_0\left[h(X)\frac{\mu_0 - X}{\sigma^2} + h'(X)\right]^2}{\mathbb{E}_0^2[h(X)]}\right) \sim \mathcal{N}\left(0, \frac{\mathbb{E}_0\left[h(X)(\mu_0 - X) + \sigma^2 h'(X)\right]^2}{\mathbb{E}_0^2[h(X)]}\right).$$

By integration by parts, (suppressing the dependence of p_{μ_0} on μ_0)

$$\begin{aligned}
&\mathbb{E}_0[h(X)h'(X)(X - \mu_0)] \\
&= \int_0^\infty h'(x)h(x)(x - \mu_0)p(x)\,\mathrm{d}x = \int_0^\infty h(x)(x - \mu_0)p(x)\,\mathrm{d}h(x) \\
&= h^2(x)(x - \mu_0)p(x)\big|_0^\infty - \int h(x)\,\mathrm{d}h(x)(x - \mu_0)p(x) \\
&= -\int h^2(x)p(x)\,\mathrm{d}x - \int h(x)h'(x)(x - \mu_0)p(x)\,\mathrm{d}x + \int h^2(x)\frac{(x - \mu_0)^2}{\sigma^2}p(x)\,\mathrm{d}x,
\end{aligned}$$

where the last step follows from the assumptions $\lim_{x \searrow 0^+} h(x) = 0$ and $\lim_{x \nearrow +\infty} h^2(x)(x - \mu_0)p_{\mu_0}(x) = 0$. So

$$\mathbb{E}_0[h(X)h'(X)(X - \mu_0)] = \frac{\mathbb{E}[h^2(X)((X - \mu_0)^2/\sigma^2 - 1)]}{2}. \tag{A.5}$$

The asymptotic variance thus becomes

$$\begin{aligned}
&\frac{\mathbb{E}_0\left[h(X)(\mu_0 - X) + \sigma^2 h'(X)\right]^2}{\mathbb{E}_0^2[h(X)]} \\
&= \frac{\mathbb{E}_0\left[h^2(X)(X - \mu_0)^2 - 2\sigma^2 h^2(X)\left((X - \mu_0)^2/\sigma^2 - 1\right)/2 + \sigma^4 h'^2(X)\right]}{\mathbb{E}_0^2[h(X)]} \\
&= \frac{\mathbb{E}_0[\sigma^2 h^2(X) + \sigma^4 h'^2(X)]}{\mathbb{E}_0^2[h(X)]}.
\end{aligned}$$

We note that the Cramér-Rao lower bound is $\frac{\sigma^4}{\operatorname{var}(X - \mu_0)}$, which follows from taking the second derivative of $\log p_{\mu_0}$ with respect to μ_0. $\square$

Proof for Example 6. For the family of univariate truncated Gaussian distributions with known mean parameter μ and unknown variance parameter $\sigma^2 > 0$, we have

$$p_\theta(x) \propto \exp\left(\theta t(x) + b(x)\right), \quad \theta \equiv \frac{1}{\sigma^2}, \quad t(x) \equiv -(x-\mu)^2/2, \quad b(x) = 0.$$

We estimate $\theta \equiv 1/\sigma^2$. By (A.1) and (A.2),

$$\begin{aligned}
\hat{\theta} &\equiv \Gamma(\boldsymbol{x})^{-1} g(\boldsymbol{x}) \\
&= -\left[\sum_{i=1}^n h(X_i) t'(X_i)^2\right]^{-1} \left[\sum_{i=1}^n h(X_i) b'(X_i) t'(X_i) + h(X_i) t''(X_i) + h'(X_i) t'(X_i)\right] \\
&= \left[\sum_{i=1}^n h(X_i)(X_i - \mu)^2\right]^{-1} \left[\sum_{i=1}^n h(X_i) + h'(X_i)(X_i - \mu)\right].
\end{aligned}$$

By Theorem 4, $\sqrt{n}(\hat{\theta} - \theta) \to_d \mathcal{N}(0, \varsigma^2)$, where

$$\begin{aligned}
\varsigma^2 &\equiv \frac{\mathbb{E}_0\left[h(X)((X-\mu)^2/\sigma_0^2 - 1) - h'(X)(X-\mu)\right]^2}{\mathbb{E}_0^2[h(X)(X-\mu)^2]} \\
&= \frac{1}{\mathbb{E}_0^2[h(X)(X-\mu)^2]} \Big(\mathbb{E}_0[h^2(X)(X-\mu)^4/\sigma_0^4 - 2h^2(X)(X-\mu)^2/\sigma_0^2 + h^2(X) + {h'}^2(X)(X-\mu)^2 \\
&\qquad - 2h(X)h'(X)(X-\mu)^3/\sigma_0^2 + 2h(X)h'(X)(X-\mu)\Big).
\end{aligned}$$

By integration by parts, (suppressing the dependence of $p_{\sigma_0^2}$ on σ_0^2)

$$\begin{aligned}
&\mathbb{E}_0[h(X)h'(X)(X-\mu)^3] \\
&= \int_0^\infty h'(x)h(x)(x-\mu)^3 p(x)\,\mathrm{d}x = \int_0^\infty h(x)(x-\mu)^3 p(x)\,\mathrm{d}h(x) \\
&= h^2(x)(x-\mu)^3 p(x)\big|_0^\infty - \int h(x)\,\mathrm{d}h(x)(x-\mu)^3 p(x) \\
&= -\int h(x)h'(x)(x-\mu)^3 p(x)\,\mathrm{d}x - 3\int h^2(x)(x-\mu)^2 p(x)\,\mathrm{d}x + \int h^2(x)\frac{(x-\mu)^4}{\sigma_0^2} p(x)\,\mathrm{d}x,
\end{aligned}$$

where the last step follows from the assumptions $\lim_{x \searrow 0^+} h(x) = 0$ and $\lim_{x \nearrow +\infty} h^2(x)(x-\mu)^3 p_{\sigma_0^2}(x) = 0$. Combining this with (A.5) we get

$$\sqrt{n}(\hat{\theta} - \theta) \to_d \mathcal{N}(0, \varsigma^2) \sim \mathcal{N}\left(0, \frac{2\mathbb{E}_0[h^2(X)(X-\mu)^2/\sigma_0^2] + \mathbb{E}_0[{h'}^2(X-\mu)^2]}{\mathbb{E}_0^2[h(X)(X-\mu)^2]}\right),$$

and so by the delta method, for $\hat{\sigma}_k^2 \equiv \hat{\theta}^{-1}$,

$$\sqrt{n}(\hat{\sigma}_h^2 - \sigma_0^2) \to_d \mathcal{N}\left(0, \frac{2\sigma_0^6 \mathbb{E}_0[h^2(X)(X-\mu)^2] + \sigma_0^8 \mathbb{E}_0[{h'}^2(X-\mu)^2]}{\mathbb{E}_0^2[h(X)(X-\mu)^2]}\right).$$

We note that the Cramér-Rao lower bound is $\frac{4\sigma_0^8}{\mathrm{var}(X-\mu)^2}$, which follows from taking the second derivative of $\log p_{\sigma_0^2}$ with respect to σ_0^2. $\square$

A.4 REGULARIZED GENERALIZED SCORE MATCHING

We first verify assumptions (A1)–(A2) in the case of truncated Gaussian distributions.

Lemma A.2 (Assumptions for truncated Gaussian)**.** *Consider the non-centered truncated Gaussian distribution with density*

$$\log p_0(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_0)^\top \mathbf{K}_0(\boldsymbol{x}-\boldsymbol{\mu}_0) + \mathrm{const}$$

with unknown positive definite inverse covariance parameter $\mathbf{K}_0$ *and unknown mean parameter* $\boldsymbol{\mu}_0$*. Then assuming* $0 \le h_j \le M_j$*,* $\lim_{x_j \searrow 0^+} h_j(x_j) = 0$ *and* $|h_j'| \le M_j'$*, assumptions (A1)–(A2) for score matching are satisfied for any proposed parameters* $\mathbf{K} \succ \mathbf{0}$ *and* $\boldsymbol{\mu}$*. Taking* $\boldsymbol{\mu} \equiv \boldsymbol{\mu}_0 \equiv \mathbf{0}$ *the assumptions also hold in the centered setting. Choosing* $m = 1$ *gives the univariate case.*

Proof of Lemma A.2. Consider $p \sim \mathrm{TN}(\boldsymbol{\mu}, \mathbf{K})$, with $\boldsymbol{k}_j$ the j-th column of $\mathbf{K}$. Let $M \equiv \max_j M_j$ and $M' \equiv \max_j M_j'$.

(A1) For any fixed $\boldsymbol{x}_{-j} \in \mathbb{R}_+^{m-1}$ and any $p \in \mathcal{P}_+$ with parameters $\mathbf{K}$ and $\boldsymbol{\mu}$,

$$\begin{aligned}\lim_{x_j \nearrow \infty} h_j(x_j)p_0(\boldsymbol{x})\partial_j \log p(\boldsymbol{x}) &\propto \lim_{x_j \nearrow \infty} h_j(x_j)\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_0)^\top \mathbf{K}_0(\boldsymbol{x}-\boldsymbol{\mu}_0)\right)\boldsymbol{k}_j^\top(\boldsymbol{x}-\boldsymbol{\mu})\\ &= \lim_{x_j \nearrow \infty} h_j(x_j)\exp\left(C_1 + C_2 x_j - \frac{1}{2}\kappa_{0,jj}x_j^2\right)(C_3 + C_4 x_j)\end{aligned}$$

for some constants C_1, C_2, C_3, and C_4 depending on $\boldsymbol{x}_{-j}$, $\mathbf{K}_0$, $\mathbf{K}$, $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}$. Since $\kappa_{0,jj} > 0$ and we assumed h_j to be bounded, the limit equals to 0 for all j and $\boldsymbol{x}_{-j}$.
Similarly,

$$\begin{aligned}\lim_{x_j \searrow 0^+} h_j(x_j)p_0(\boldsymbol{x})\partial_j \log p(\boldsymbol{x}) &\propto \lim_{x_j \searrow 0^+} h_j(x_j)\exp\left(C_1 + C_2 x_j - \frac{1}{2}\kappa_{0,jj}x_j^2\right)(C_3 + C_4 x_j)\\ &= \exp(C_1)C_3 \lim_{x_j \searrow 0^+} h_j(x_j) = 0\end{aligned}$$

if and only if we assume $\lim_{x_j \searrow 0^+} h_j(x_j) = 0$.

(A2) For any $p \in \mathcal{P}_+$ with parameters $\mathbf{K}$ and $\boldsymbol{\mu}$,

$$\begin{aligned}\mathbb{E}_{p_0}\|\nabla \log p(\boldsymbol{X}) \circ \boldsymbol{h}^{1/2}(\boldsymbol{X})\|_2^2 \le M\mathbb{E}_{p_0}\|\nabla \log p(\boldsymbol{X})\|_2^2 &= M\mathrm{tr}\left(\mathbb{E}_{p_0}\left[(\mathbf{K}(\boldsymbol{X}-\boldsymbol{\mu}))(\mathbf{K}(\boldsymbol{X}-\boldsymbol{\mu}))^\top\right]\right)\\ &= M\mathrm{tr}\left(\mathbf{K}\mathbb{E}_{p_0}\left[(\boldsymbol{X}-\boldsymbol{\mu}_0+(\boldsymbol{\mu}_0-\boldsymbol{\mu}))(\boldsymbol{X}-\boldsymbol{\mu}_0+(\boldsymbol{\mu}_0-\boldsymbol{\mu}))^\top\right]\mathbf{K}^\top\right)\\ &= M\mathrm{tr}\left(\mathbf{K}\left(\mathbf{K}_0^{-1}+(\boldsymbol{\mu}_0-\boldsymbol{\mu})(\boldsymbol{\mu}_0-\boldsymbol{\mu})^\top\right)\mathbf{K}\right) < +\infty\end{aligned}$$

since M, $\mathbf{K}$, $\mathbf{K}_0$, $\boldsymbol{\mu}$, $\boldsymbol{\mu}_0$ are all finite constants. We also have

$$\begin{aligned}\mathbb{E}_{p_0}\|(\nabla \log p(\boldsymbol{X}) \circ \boldsymbol{h}(\boldsymbol{X}))'\|_1 &= \sum_{i=1}^m \mathbb{E}_{p_0}\left|h_j'(X_j)\partial_j \log p(\boldsymbol{X}) + h_j(X_j)\partial_j^2 \log p(\boldsymbol{X})\right|\\ &\le \sum_{i=1}^m \mathbb{E}_{p_0}|h_j'(X_j)\partial_j \log p(\boldsymbol{X})| + \mathbb{E}_{p_0}|h_j(X_j)\partial_j^2 \log p(\boldsymbol{X})|\\ &\le \sum_{i=1}^m M'\mathbb{E}_{p_0}|\boldsymbol{k}_j^\top(\boldsymbol{X}-\boldsymbol{\mu})| + M\kappa_{jj}\\ &\le \sum_{i=1}^m M'|\boldsymbol{k}_j|^\top \mathbb{E}_{p_0}\boldsymbol{X} + M'|\boldsymbol{k}_j^\top \boldsymbol{\mu}| + M\mathrm{tr}(\mathbf{K}) < +\infty.\end{aligned}$$

Hence, (A1) and (A2) are both satisfied.

□

Our analysis of the regularized generalized $\boldsymbol{h}$-score matching estimator follows the proof for the following theorem from Lin et al. (2016), restated below. In our definition and implementation we choose to optimize over all symmetric matrices, but we adopt the following theorem in whose proof the symmetry condition is not explicitly imposed, in order to decouple the columns of $\mathbf{K}$ and to highlight the scaling.

Theorem A.3 (Analog of Theorem 1 from Lin et al. (2016))**.** *Recall that* $S_0 \equiv S(\mathbf{K}_0) \equiv \{(i,j) : \kappa_{0,ij} \neq 0\}$. *Suppose* $\mathbf{\Gamma}_{0,S_0S_0}$ *is invertible and satisfies the irrepresentability condition (10) with incoherence parameter* $\alpha \in (0,1]$. *Assume that*

$$\|\mathbf{\Gamma}(\mathbf{x}) - \mathbf{\Gamma}_0\|_\infty < \epsilon_1, \qquad \|\boldsymbol{g}(\mathbf{x}) - \boldsymbol{g}_0\|_\infty < \epsilon_2, \tag{A.6}$$

with $d_{\mathbf{K}_0}\epsilon_1 \leq \alpha/(6c_{\mathbf{\Gamma}_0})$. *If*

$$\lambda > \frac{3(2-\alpha)}{\alpha}\max\{c_{\mathbf{K}_0}\epsilon_1, \epsilon_2\},$$

then the following statements hold:

(a) The regularized generalized $\boldsymbol{h}$*-score matching estimator* $\hat{\mathbf{K}}$ *in (9) is unique, with support* $\hat{S} \equiv S(\hat{\mathbf{K}}) \subseteq S_0$, *and satisfies*

$$\|\hat{\mathbf{K}} - \mathbf{K}_0\|_\infty \leq \frac{c_{\mathbf{\Gamma}_0}}{2-\alpha}\lambda.$$

(b) If

$$\min_{1\leq j<k\leq m} |\mathbf{K}_{0,jk}| > \frac{c_{\mathbf{\Gamma}_0}}{2-\alpha}\lambda,$$

then $\hat{S} = S_0$ *and* $\operatorname{sign}(\hat{\mathbf{K}}_{jk}) = \operatorname{sign}(\mathbf{K}_{0.jk})$ *for all* $(j,k) \in S_0$.

This is a deterministic result, and the improvement of our generalized estimator over the one in Lin et al. (2016) is in its asymptotic guarantees, as in Theorem 10. We present a corollary to this theorem, as seen in the second and third inequalities in Theorem 10 (a).

Corollary A.1. *Suppose the same assumptions under Theorem A.3 hold. Then* $\hat{\mathbf{K}}$ *satisfies*

$$\begin{aligned}
\|\hat{\mathbf{K}} - \mathbf{K}_0\|_F &\leq \frac{c_{\mathbf{\Gamma}_0}}{2-\alpha}\lambda\sqrt{|S_0|} \leq \frac{c_{\mathbf{\Gamma}_0}}{2-\alpha}\lambda\sqrt{d_{\mathbf{K}_0}m},\\
\|\hat{\mathbf{K}} - \mathbf{K}_0\|_2 &\leq \frac{c_{\mathbf{\Gamma}_0}}{2-\alpha}\lambda\min(\sqrt{|S_0|}, d_{\mathbf{K}_0}).
\end{aligned}$$

Proof of Corollary A.1. By Theorem A.3, under assumptions in that theorem, the support of $\hat{\mathbf{K}}$ is a subset of the true support of $\mathbf{K}_0$, and $\|\hat{\mathbf{K}} - \mathbf{K}_0\|_\infty \leq \frac{c_{\mathbf{\Gamma}_0}}{2-\alpha}\lambda$. Since $\mathbf{K}_0$ has $|S_0|$ nonzero entries,

$$|||\hat{\mathbf{K}} - \mathbf{K}_0|||_F = \left[\sum_{\mathbf{K}_{0,jk}\neq 0}(\hat{\mathbf{K}}_{jk} - \mathbf{K}_{0,jk})^2\right]^{1/2} \leq \sqrt{|S_0|}\|\hat{\mathbf{K}} - \mathbf{K}_0\|_\infty \leq \frac{c_{\mathbf{\Gamma}_0}}{2-\alpha}\lambda\sqrt{|S_0|}.$$

Similarly, by the definition of matrix ℓ_∞-ℓ_∞ norm,

$$|||\hat{\mathbf{K}} - \mathbf{K}_0|||_2 \leq |||\hat{\mathbf{K}} - \mathbf{K}_0|||_\infty = \max_{j=1,\ldots,m}\sum_{k=1}^m |\hat{\mathbf{K}}_{jk} - \mathbf{K}_{0,jk}| \leq \frac{c_{\mathbf{\Gamma}_0}}{2-\alpha}\lambda d_{\mathbf{K}_0}.$$

The result follows by also noting that $|||\hat{\mathbf{K}} - \mathbf{K}_0|||_2 \leq |||\hat{\mathbf{K}} - \mathbf{K}_0|||_F$. □

Proof of Theorem 10. By Theorem A.3 it suffices to prove that for any $\tau > 3$, we can bound $\|\mathbf{\Gamma}(\mathbf{x}) - \mathbf{\Gamma}_0\|_\infty$ by some ϵ_1 and $\|\boldsymbol{g}(\mathbf{x}) - \boldsymbol{g}_0\|_\infty$ by some ϵ_2, uniformly with probability $1 - m^{3-\tau}$. Recall from Section 4.2 that the j^{th} block of $\mathbf{\Gamma} \in \mathbb{R}^{m^2\times m^2}$ has (k,ℓ)-th entry

$$\frac{1}{n}\sum_{i=1}^n X_k^{(i)}X_\ell^{(i)}h_j(X_j^{(i)}),$$

and the entry in $\boldsymbol{g} \in \mathbb{R}^{m^2}$ (obtained by linearizing a $m \times m$ matrix) corresponding to (j,k) with $j \neq k$, is

$$\frac{1}{n}\sum_{i=1}^n X_k^{(i)}h_j'(X_j^{(i)}),$$

while the entry for (j,j) is

$$\frac{1}{n}\sum_{i=1}^{n} X_j^{(i)} h_j'(X_j^{(i)}) + \frac{1}{n}\sum_{i=1}^{n} h_j(X_j^{(i)}).$$

Denote $M \equiv \max_j \sup_{x>0} h_j(x)$ and $M' \equiv \max_j \sup_{x>0} h_j'(x)$, and let $c_{\boldsymbol{X}} \equiv 2\max_j(2\sqrt{\Sigma_{jj}} + \sqrt{e}\mathbb{E}_0 X_j)$. Using results for sub-gaussian random variables from Lemma A.6 below and Hoeffding's inequality, we have for any $t_1, t_{2,1}, t_{2,2} > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_k^{(i)} X_\ell^{(i)} h_j(X_j^{(i)}) - \mathbb{E}_0 X_k X_\ell h_j(X_j)\right| > t_1\right) \leq 2\exp\left(-\min\left(\frac{nt_1^2}{2M^2 c_{\boldsymbol{X}}^4}, \frac{nt_1}{2Mc_{\boldsymbol{X}}^2}\right)\right),$$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_k^{(i)} h_j'(X_j^{(i)}) - \mathbb{E}_0 X_k h_j'(X_j)\right| \geq t_{2,1}\right) \leq 2\exp\left(-\frac{nt_{2,1}^2}{2M'^2 c_{\boldsymbol{X}}^2}\right),$$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} h_j(X_j^{(i)}) - \mathbb{E}_0 h_j(X_j)\right| \geq t_{2,2}\right) \leq 2\exp\left(-2nt_{2,2}^2/M^2\right).$$

Choosing

$$\epsilon_1 \equiv Mc_{\boldsymbol{X}}^2 \max\left\{\frac{2(\log m^\tau + \log 6)}{n}, \sqrt{\frac{2(\log m^\tau + \log 6)}{n}}\right\},$$

$$\epsilon_{2,1} \equiv \sqrt{2}M'c_{\boldsymbol{X}}\sqrt{\frac{\log m^{\tau-1} + \log 6}{n}}, \quad \epsilon_{2,2} \equiv M\sqrt{\frac{\log m^{\tau-2} + \log 6}{2n}},$$

and taking union bounds over m^3, m^2, and m events, respectively, we have

$$\mathbb{P}\left(\sup_{j,k,\ell}\left|\frac{1}{n}\sum_{i=1}^{n} X_k^{(i)} X_\ell^{(i)} h_j(X_j^{(i)}) - \mathbb{E}_0 X_k X_\ell h_j(X_j)\right| \geq \epsilon_1\right) \leq \frac{1}{3m^{\tau-3}},$$

$$\mathbb{P}\left(\sup_{j,k}\left|\frac{1}{n}\sum_{i=1}^{n} X_k^{(i)} h_j'(X_j^{(i)}) - \mathbb{E}_0 X_k h_j'(X_j)\right| \geq \epsilon_{2,1}\right) \leq \frac{1}{3m^{\tau-3}},$$

$$\mathbb{P}\left(\sup_{j}\left|\frac{1}{n}\sum_{i=1}^{n} h_j(X_j^{(i)}) - \mathbb{E}_0 h_j(X_j)\right| \geq \epsilon_{2,2}\right) \leq \frac{1}{3m^{\tau-3}}.$$

Hence, with probability at least $1 - m^{3-\tau}$, $\|\boldsymbol{\Gamma}(\mathbf{x}) - \boldsymbol{\Gamma}_0\|_\infty < \epsilon_1$ and $\|\boldsymbol{g}(\mathbf{x}) - \boldsymbol{g}_0\|_\infty < \epsilon_2 \equiv \epsilon_{2,1} + \epsilon_{2,2}$. Consider any $\tau > 3$, and let

$$\begin{aligned}
c_2 &\equiv \frac{6}{\alpha}c_{\boldsymbol{\Gamma}_0}, \qquad n \geq \max\{2M^2 c_{\boldsymbol{X}}^4 c_2^2 d_{\mathbf{K}_0}^2(\tau\log m + \log 6), 2Mc_{\boldsymbol{X}}^2 c_2 d_{\mathbf{K}_0}(\tau\log m + \log 6)\},\\
\lambda &> \frac{3(2-\alpha)}{\alpha}\max\{c_{\mathbf{K}_0}\epsilon_1, \epsilon_2\}\\
&\equiv \frac{3(2-\alpha)}{\alpha}\max\left\{Mc_{\mathbf{K}_0}c_{\boldsymbol{X}}^2\frac{2(\log m^\tau + \log 6)}{n},\right.\\
&\qquad \left. Mc_{\mathbf{K}_0}c_{\boldsymbol{X}}^2\sqrt{\frac{2(\log m^\tau + \log 6)}{n}}, \sqrt{2}M'c_{\boldsymbol{X}}\sqrt{\frac{\log m^{\tau-1} + \log 6}{n}} + M\sqrt{\frac{\log m^{\tau-2} + \log 6}{2n}}\right\}.
\end{aligned}$$

Then $d_{\mathbf{K}_0}\epsilon_1 \leq \alpha/(6c_{\boldsymbol{\Gamma}_0})$ and the results follow from Theorem A.3. □

We now present the definition of sub-Gaussian and sub-exponential norms and variables as well as lemmas required for the proof above.

Definition A.4 (Sub-Gaussian and Sub-Exponential Variables)**.** The *sub-gaussian* ($r=2$) and *sub-exponential* ($r=1$) *norms* of a random variable are defined as

$$\|X\|_{\psi_r} \equiv \sup_{q\geq 1} q^{-1/r}(\mathbb{E}|X|^{rq})^{1/(rq)} \equiv \sup_{q\geq 1} q^{-1/r}\|X\|_{rq}.$$

If $\|X\|_{\psi_2} < \infty$ we say X is *sub-gaussian*; if $\|X\|_{\psi_1} < \infty$ we call X *sub-exponential*.
For a *zero-mean* sub-gaussian random variable X also define the *sub-gaussian parameter*

$$\tau(X) = \inf\{\tau \geq 0 : \mathbb{E}\exp(tX) \leq \exp(\tau^2 t^2/2),\ \forall t \in \mathbb{R}\}.$$

Note that the definition of sub-gaussian norm here allows the variable to be non-centered, and is different from the one in Vershynin (2010), which uses $\|X\|_q$ in the definition. Instead, it coincides with θ_2 in Buldygin and Kozachenko (2000). The definition of the sub-gaussian parameter is the same as in Buldygin and Kozachenko (2000), and the definition of the sub-exponential norm is as in Vershynin (2010).

Lemma A.5 (Properties of Sub-Gaussian and Sub-Exponential Variables)**.** *Then*

1) *For any X and $r = 1, 2$, $\|X - \mathbb{E}X\|_{\psi_r} \leq 2\|X\|_{\psi_r}$ and $\|X\|_{\psi_r} \leq \|X - \mathbb{E}X\|_{\psi_r} + |\mathbb{E}X|$, as long as the expectation and norms are finite.*

2) *(Buldygin and Kozachenko, 2000) $\tau(X)$ is a norm on the space of all zero-mean sub-gaussian variables; in particular, $\tau(X + Y) \leq \tau(X) + \tau(Y)$ as long as the quantities are defined and finite.*
If X is zero-mean sub-gaussian, then $\mathrm{var}(X) \leq \tau^2(X)$, $\|X\|_{\psi_2} \leq 2\tau(X)/\sqrt{e}$, $\tau(X) \leq \sqrt{e}\|X\|_{\psi_2}$.
If $X_1, \ldots, X_n$ are i.i.d. zero-mean sub-gaussian, $\tau\left(\frac{1}{n}\sum_{i=1}^n X_i\right) \leq \frac{1}{\sqrt{n}}\tau(X_i)$.

3) *If random variables X_1 and X_2 (not necessarily independent) are sub-gaussian with $\|X_1\|_{\psi_2} \leq K_1$ and $\|X_2\|_{\psi_2} \leq K_2$, then X_1X_2 is sub-exponential with $\|X_1X_2\|_{\psi_1} \leq K_1K_2$.*

4) *(Buldygin and Kozachenko, 2000) If X is zero-mean sub-gaussian,*

$$\mathbb{E}|X|^q \leq 2(q/e)^{q/2}\tau^q(X)$$

for any $q > 0$.

5) *(Buldygin and Kozachenko, 2000) If $X_1, \ldots, X_n$ are independent zero-mean sub-gaussian variables, then for any $\epsilon > 0$,*

$$\mathbb{P}(|X_1| \geq \epsilon) \leq 2\exp\left(-\frac{\epsilon^2}{2\tau^2(X_1)}\right),$$
$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i\right| > \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{2\max_i \tau^2(X_i)}\right).$$

6) *(Vershynin, 2010) If $X_1, \ldots, X_n$ are independent zero-mean sub-exponential random variables with $K \geq \max_i \|X_i\|_{\psi_1}$, then for any $\epsilon > 0$,*

$$\mathbb{P}(|X_1| \geq \epsilon) \leq 2\exp\left(-\min\left(\frac{\epsilon^2}{8e^2K^2}, \frac{\epsilon}{4eK}\right)\right),$$
$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i\right| \geq \epsilon\right) \leq 2\exp\left(-\min\left(\frac{n\epsilon^2}{8e^2K^2}, \frac{n\epsilon}{4eK}\right)\right).$$

Proof. 1) For $r = 1, 2$, by triangle inequality, $\|X - \mathbb{E}X\|_{\psi_r} \leq \|X\|_{\psi_r} + \|\mathbb{E}X\|_{\psi_r} = \|X\|_{\psi_r} + |\mathbb{E}X| \leq \|X\|_{\psi_r} + \mathbb{E}|X| \leq 2\|X\|_{\psi_r}$, where in the last step we used the definition of $\|\cdot\|_{\psi_r}$ with $q = 1$ for $r = 1$ and $\mathbb{E}|X| \leq (\mathbb{E}|X|^2)^{1/2}$ with $q = 2$ for $r = 2$. On the other hand, $\|X\|_{\psi_r} \leq \|X - \mathbb{E}X\|_{\psi_r} + \|\mathbb{E}X\|_{\psi_r} = \|X - \mathbb{E}X\|_{\psi_r} + |\mathbb{E}X|$.

2) These follow from Theorems 1.2 and 1.3 and Lemmas 1.2 and 1.7 from Buldygin and Kozachenko (2000), and $\sqrt[4]{3.1}e^{9/16}/\sqrt{2} \approx 1.6467 \leq 1.6487 \approx \sqrt{e}$.

3) By Hölder's inequality (or Cauchy-Schwarz),

$$\begin{aligned}
\|X_1X_2\|_{\psi_1} &= \sup_{q\geq 1} q^{-1}(\mathbb{E}|X_1X_2|^q)^{1/q} = \sup_{q\geq 1} q^{-1}(\mathbb{E}|X_1^qX_2^q|)^{1/q} \\
&\leq \sup_{q\geq 1} q^{-1}\left[(\mathbb{E}|X_1|^{2q})^{1/2}(\mathbb{E}|X_2|^{2q})^{1/2}\right]^{1/q} \\
&\leq \sup_{q\geq 1}\left[q^{-1/2}(\mathbb{E}|X_1|^{2q})^{1/2q}\right]\sup_{q\geq 1}\left[q^{-1/2}(\mathbb{E}|X_2|^{2q})^{1/2q}\right] \\
&= \|X_1\|_{\psi_2}\|X_2\|_{\psi_2} \leq K_1K_2.
\end{aligned}$$

4) This is Lemma 1.4 from Buldygin and Kozachenko (2000).

5) This is Theorem 1.5 from Buldygin and Kozachenko (2000).

6) This follows from Corollary 5.17 from Vershynin (2010). $\square$

Lemma A.6. *Suppose $\boldsymbol{X}$ follows a truncated normal distribution on $\mathbb{R}_+^m$ with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma} = \mathbf{K}^{-1} \succ \mathbf{0}$. Let $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(n)}$ be i.i.d. copies of $\boldsymbol{X}$, with j-th component of the i-th copy being $X_j^{(i)}$. Then*

1. *For $j = 1, \ldots, p$, $\tau(X_j - \mathbb{E}X_j) \leq \sqrt{\Sigma_{jj}}$. That is, the sub-gaussian parameter of any marginal distribution of $\boldsymbol{X}$, after centering, is bounded by the square root of its corresponding diagonal entry in the covariance parameter $\boldsymbol{\Sigma}$. Then for any $\epsilon > 0$,*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_j^{(i)} - \mathbb{E}X_j\right| > \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{2\Sigma_{jj}}\right).$$

In particular, if h_0 is a function bounded by M_0, then for any $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_j^{(i)} h_0(X_k^{(i)}) - \mathbb{E}X_j h_0(X_k)\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{8M_0^2(2\sqrt{\Sigma_{jj}} + \sqrt{e}\mathbb{E}X_j)^2}\right),$$
$$\tau\left(\frac{1}{n}\sum_{i=1}^n X_j^{(i)} h_0(X_k^{(i)}) - \mathbb{E}X_j h_0(X_k)\right) \leq \frac{2M_0}{\sqrt{n}}(2\sqrt{\Sigma_{jj}} + \sqrt{e}\mathbb{E}X_j),$$
$$\left\|\frac{1}{n}\sum_{i=1}^n X_j^{(i)} h_0(X_k^{(i)}) - \mathbb{E}X_j h_0(X_k)\right\|_{\psi_2} \leq \frac{4M_0}{\sqrt{en}}(2\sqrt{\Sigma_{jj}} + \sqrt{e}\mathbb{E}X_j).$$

2. *For $j, k, \ell \in \{1, \ldots, p\}$, if h_0 is a function bounded by M_0, then with $c_{\boldsymbol{X}} \equiv 2\max_j(2\sqrt{\Sigma_{jj}} + \sqrt{e}\mathbb{E}X_j)$,*

$$\|X_j X_k h_0(X_\ell) - \mathbb{E}X_j X_k h_0(X_\ell)\|_{\psi_1} \leq \frac{M_0}{2e} c_{\boldsymbol{X}}^2. \tag{A.7}$$

In particular, for any $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_j^{(i)} X_k^{(i)} h_0(X_\ell^{(i)}) - \mathbb{E}X_j X_k h_0(X_\ell)\right| > \epsilon\right) \leq 2\exp\left(-\min\left(\frac{n\epsilon^2}{2M_0^2 c_{\boldsymbol{X}}^4}, \frac{n\epsilon}{2M_0 c_{\boldsymbol{X}}^2}\right)\right).$$

Proof of Lemma A.6. 1. Without loss of generality choose $j = 1$. By the definition of sub-gaussian parameters, we need to show that for all $t \in \mathbb{R}$,

$$\mathbb{E}\exp(tX_1) \leq \exp(t^2\Sigma_{11}/2 + t\mathbb{E}X_1),$$

which is equivalent to

$$t^2\Sigma_{11}/2 + t\mathbb{E}X_1 - \log\mathbb{E}\exp(tX_1) \geq 0 \quad \forall t \in \mathbb{R}. \tag{A.8}$$

Since the left-hand side of (A.8) equals 0 at $t = 0$, it suffices to show that its derivative

$$t\Sigma_{11} + \mathbb{E}X_1 - \frac{\mathrm{d}\log\mathbb{E}\exp(tX_1)}{\mathrm{d}t} = t\Sigma_{11} + \mathbb{E}X_1 - \frac{\frac{\mathrm{d}\mathbb{E}\exp(tX_1)}{\mathrm{d}t}}{\mathbb{E}\exp(tX_1)} \tag{A.9}$$

is non-negative on $(0, \infty)$ and non-positive on $(-\infty, 0)$. By properties of moment-generating functions, $\frac{\mathrm{d}\mathbb{E}\exp(tX_1)}{\mathrm{d}t}$ evaluated at $t = 0$ equals $\mathbb{E}X_1$, so (A.9) equals 0 at $t = 0$. It in turn suffices to show the derivative of (A.9), namely

$$\Sigma_{11} - \frac{\mathrm{d}^2\log\mathbb{E}\exp(tX_1)}{\mathrm{d}t^2} \tag{A.10}$$

is non-negative in $t \in \mathbb{R}$.

By Tallis (1961), denoting the first column of $\mathbf{\Sigma}$ as $\mathbf{\Sigma}_1$, the moment-generating function of the marginal distribution of X_1 is

$$\frac{\int_{\mathbb{R}_+^p-\boldsymbol{\mu}-t\mathbf{\Sigma}_1}\exp\left(-\frac{1}{2}\boldsymbol{x}^\top\mathbf{\Sigma}^{-1}\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}}{\int_{\mathbb{R}_+^p-\boldsymbol{\mu}}\exp\left(-\frac{1}{2}\boldsymbol{x}^\top\mathbf{\Sigma}^{-1}\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}}\exp\left(t\mu_1+\frac{1}{2}t^2\Sigma_{11}^2\right).$$

(A.10) thus becomes

$$-\frac{\mathrm{d}^2}{\mathrm{d}t^2}\log\int_{\mathbb{R}_+^p-\boldsymbol{\mu}-t\mathbf{\Sigma}_1}\exp\left(-\frac{1}{2}\boldsymbol{x}^\top\mathbf{\Sigma}^{-1}\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}.$$

Showing this is non-negative in $t\in\mathbb{R}$ is equivalent to showing that the integral itself is log-concave in t. But

$$\int_{\mathbb{R}_+^p-\boldsymbol{\mu}-t\mathbf{\Sigma}_1}\exp\left(-\frac{1}{2}\boldsymbol{x}^\top\mathbf{\Sigma}^{-1}\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}=\int_{\mathbb{R}^p}\exp\left(-\frac{1}{2}\boldsymbol{x}^\top\mathbf{\Sigma}^{-1}\boldsymbol{x}\right)\mathbb{1}_{\mathbb{R}_+^p-\boldsymbol{\mu}}(\boldsymbol{x}+t\mathbf{\Sigma}_1)\,\mathrm{d}\boldsymbol{x}$$

with $\exp\left(-\frac{1}{2}\boldsymbol{x}^\top\mathbf{\Sigma}^{-1}\boldsymbol{x}\right)$ log-concave in $\boldsymbol{x}$ and $\mathbb{1}_{\mathbb{R}_+^p-\boldsymbol{\mu}}(\boldsymbol{x}+t\mathbf{\Sigma}_1)$ log-concave in $(\boldsymbol{x},t)$ since $\mathbb{R}_+^p-\boldsymbol{\mu}$ is a convex set (half-space). Here $\mathbb{1}_S(\cdot)$ is the indicator function of a set S. Since log-concavity is closed under multiplication and integration over $\mathbb{R}^p$, the integral is indeed log-concave, and our proof of the bound on the sub-gaussian parameter of $X_j-\mathbb{E}X_j$ is complete. The tail bound follows from 5) of Lemma A.5.

Now by 1) and 2) of Lemma A.5,

$$\|X_j\|_{\psi_2}\le 2\sqrt{\Sigma_{jj}/e}+\mathbb{E}X_j.$$

If h_0 is a function bounded by M_0, then by definition

$$\|X_jh_0(X_k)\|_{\psi_2}\le M_0\left(2\sqrt{\Sigma_{jj}/e}+\mathbb{E}X_j\right).$$

By 1) and 2) of Lemma A.5 again,

$$\begin{aligned}\tau(X_jh_0(X_k)-\mathbb{E}X_jh_0(X_k))&\le\sqrt{e}\|X_jh_0(X_k)-\mathbb{E}X_jh_0(X_k)\|_{\psi_2}\\&\le 2\sqrt{e}\|X_jh_0(X_k)\|_{\psi_2}\\&\le 2M_0(2\sqrt{\Sigma_{jj}}+\sqrt{e}\mathbb{E}X_j).\end{aligned}$$

The tail bound thus follows from the first inequality using 5) of Lemma A.5. By 2),

$$\begin{aligned}\tau\left(\frac{1}{n}\sum_{i=1}^n X_j^{(i)}h_0(X_k^{(i)})-\mathbb{E}X_jh_0(X_k)\right)&\le\frac{2M_0}{\sqrt{n}}(2\sqrt{\Sigma_{jj}}+\sqrt{e}\mathbb{E}X_j),\\\left\|\frac{1}{n}\sum_{i=1}^n X_j^{(i)}h_0(X_k^{(i)})-\mathbb{E}X_jh_0(X_k)\right\|_{\psi_2}&\le\frac{4M_0}{\sqrt{en}}(2\sqrt{\Sigma_{jj}}+\sqrt{e}\mathbb{E}X_j).\end{aligned}$$

2. By the proof of 1) of this lemma, $\|X_j\|_{\psi_2}\le 2\sqrt{\Sigma_{jj}/e}+\mathbb{E}X_j$, and by 3) of Lemma A.5,

$$\|X_jX_k\|_{\psi_1}\le(2\sqrt{\Sigma_{jj}/e}+\mathbb{E}X_j)(2\sqrt{\Sigma_{kk}/e}+\mathbb{E}X_k)\le\max_j\left(2\sqrt{\Sigma_{jj}/e}+\mathbb{E}X_j\right)^2.$$

Since h_0 is a function bounded by M_0, by definition

$$\|X_jX_kh_0(X_\ell)\|_{\psi_1}\le M_0\max_j\left(2\sqrt{\Sigma_{jj}/e}+\mathbb{E}X_j\right)^2.$$

Then by 1) of Lemma A.5 again,

$$\|X_jX_kh_0(X_\ell)-\mathbb{E}X_jX_kh_0(X_\ell)\|_{\psi_1}\le 2M_0\max_j\left(2\sqrt{\Sigma_{jj}/e}+\mathbb{E}X_j\right)^2.$$

The tail bound then follows from 6) of Lemma A.5.

□