Supplementary material for "Minimax quasi-Bayesian estimation in sparse canonical correlation analysis via a Rayleigh quotient function"

Qiuyun Zhu and Yves Atchade
Department of Mathematics and Statistics, Boston University

S-1 Proofs

Throughout the proofs c_0 denotes a generic absolute constant that depends only on $\underline{\kappa}$ and $\bar{\kappa}$ in H3, but whose actual value or expression may change during the text. From the definition of $\Pi(\cdot|\mathbf{Z})$, for any measurable subset C of $\Delta_s \times \mathbb{R}^p$, by integrating out the non-selected component $\theta - \theta_\delta$, we have

$$\begin{aligned}\Pi(C|\mathbf{Z}) &= \frac{\sum_{\delta\in\Delta_s} e^{\mathsf{a}\|\delta\|_0} \int_{\mathbb{R}^p} \mathbf{1}_C(\delta,\theta)\,(\theta) \exp\left(-\frac{\rho_1}{2}\|\theta_\delta\|_2^2 - \frac{\rho_0}{2}\|\theta-\theta_\delta\|_2^2 + \sigma_n \mathsf{R}_n(\theta_\delta;\mathbf{Z})\right) \mathrm{d}\theta}{\sum_{\delta\in\Delta_s} e^{\mathsf{a}\|\delta\|_0} \int_{\mathbb{R}^p} \exp\left(-\frac{\rho_1}{2}\|\theta_\delta\|_2^2 - \frac{\rho_0}{2}\|\theta-\theta_\delta\|_2^2 + \sigma_n \mathsf{R}_n(\theta_\delta;\mathbf{Z})\right) \mathrm{d}\theta} \\ &= \frac{\sum_{\delta\in\Delta_s} \left(\frac{1}{p^{\mathsf{u}}}\sqrt{\frac{\rho_1}{2\pi}}\right)^{\|\delta\|_0} \int_{\mathbb{R}^{\|\delta\|_0}} \mathbf{1}_C(\delta,(u,0)_\delta) \exp\left(-\frac{\rho_1}{2}\|u\|_2^2 + \sigma_n \bar{\mathsf{R}}_n((u,0)_\delta;\mathbf{Z})\right) \mathrm{d}u}{\sum_{\delta\in\Delta_s} \left(\frac{1}{p^{\mathsf{u}}}\sqrt{\frac{\rho_1}{2\pi}}\right)^{\|\delta\|_0} \int_{\mathbb{R}^{\|\delta\|_0}} \exp\left(-\frac{\rho_1}{2}\|u\|_2^2 + \sigma_n \bar{\mathsf{R}}_n((u,0)_\delta;\mathbf{Z})\right) \mathrm{d}u}, \quad \text{(S-1)}\end{aligned}$$

where

$$\bar{\mathsf{R}}_n(\theta;\mathbf{Z}) \stackrel{\text{def}}{=} \mathsf{R}_n(\theta;\mathbf{Z}) - \mathsf{R}_n(\theta_\star;\mathbf{Z}).$$

S-1.1 Proof of Theorem 2

We recall that $\Delta_s \stackrel{\text{def}}{=} \{\delta \in \Delta : \ \|\delta\|_0 \leq s\}$, and for $\delta \in \Delta_s$, we let

$$\mathsf{B}_\delta \stackrel{\text{def}}{=} \left\{ \theta \in \mathbb{R}^p : \ \left\| \frac{\theta_\delta \theta_\delta^{\mathrm{T}}}{\|\theta_\delta\|_2^2} - \theta_\star \theta_\star^{\mathrm{T}} \right\|_{\mathsf{F}} \leq M\epsilon \right\},$$

and B_δ^c its complement in $\mathbb{R}^p$. We then set

$$\mathsf{B} \stackrel{\text{def}}{=} \bigcup_{\delta \in \Delta_s} \{\delta\} \times \mathsf{B}_\delta, \quad \text{and} \quad \mathsf{A} \stackrel{\text{def}}{=} \bigcup_{\delta \in \Delta_s} \{\delta\} \times \mathsf{B}_\delta^c.$$

Clearly we have $\Delta_s \times \mathbb{R}^p = \mathsf{B} \cup \mathsf{A}$. Hence our objective is to establish that $\Pi(\mathsf{A}|\mathbf{Z})$ is small. We show in Lemma S-1 that the denominator on the right hand side of (S-1) is bound from below by

$$\varpi \stackrel{\text{def}}{=} e^{-s_\star(\mathsf{u}+1)\log(p)}.$$

Equation (S-1) then implies that

$$\Pi(\mathsf{A}_1|\mathbf{Z}) \leq \frac{1}{\varpi} \times \sum_{\delta \in \Delta_s} \left(\frac{1}{p^{\mathsf{u}}} \sqrt{\frac{\rho_1}{2\pi}} \right)^{\|\delta\|_0} \int_{\mathbb{R}^{\|\delta\|_0}} \mathbf{1}_{\mathsf{B}_\delta^c}((u,0)_\delta) \exp\left(-\frac{\rho_1}{2}\|u\|_2^2 + \sigma_n \bar{\mathsf{R}}_n((u,0)_\delta; \mathbf{Z}) \right) \mathrm{d}u. \quad \text{(S-2)}$$

We show in Lemma S-4 that any $\theta \in \mathbb{R}^p$, such that $\|\theta\|_0 \leq s$,

$$\mathsf{R}_n(\theta; \mathbf{Z}) - \mathsf{R}_n(\theta_\star; \mathbf{Z}) \leq -\frac{\mathsf{gap}}{2} \left(\frac{\underline{\kappa}}{\bar{\kappa}} \right)^2 \|\theta\theta^{\mathrm{T}} - \theta_\star \theta_\star^{\mathrm{T}}\|_{\mathsf{F}}^2 + c_0 \mathsf{r}_1 \|\theta\theta^{\mathrm{T}} - \theta_\star \theta_\star^{\mathrm{T}}\|_{\mathsf{F}},$$

for some absolute constant c_0 that depends only on $\underline{\kappa}$ and $\bar{\kappa}$. Therefore, for $\frac{4c_0\mathsf{r}_1}{\mathsf{gap}} \left(\frac{\bar{\kappa}}{\underline{\kappa}} \right)^2 \leq \|\theta\theta^{\mathrm{T}} - \theta_\star \theta_\star^{\mathrm{T}}\|_{\mathsf{F}}$, we have

$$\mathsf{R}_n(\theta; \mathbf{Z}) - \mathsf{R}_n(\theta_\star; \mathbf{Z}) \leq -\frac{\mathsf{gap}}{4} \left(\frac{\underline{\kappa}}{\bar{\kappa}} \right)^2 \|\theta\theta^{\mathrm{T}} - \theta_\star \theta_\star^{\mathrm{T}}\|_{\mathsf{F}}^2.$$

Therefore, for $M \geq 4c_0 \left(\frac{\bar{\kappa}}{\underline{\kappa}}\right)^2$, (S-2) becomes

$$\begin{aligned}\Pi(\mathsf{A}_1|\mathbf{Z}) &\leq \frac{1}{\varpi} e^{-\frac{M^2 \mathsf{gap}}{4}\left(\frac{\underline{\kappa}}{\bar{\kappa}}\right)^2 \sigma_n \epsilon^2} \sum_{\delta \in \Delta_s} \left(\frac{1}{p^{\mathsf{u}}}\sqrt{\frac{\rho_1}{2\pi}}\right)^{\|\delta\|_0} \int_{\mathbb{R}^{\|\delta\|_0}} \mathbf{1}_{\mathsf{B}_\delta}((u,0)_\delta) \exp\left(-\frac{\rho_1}{2}\|u\|_2^2\right) \mathrm{d}u \\ &\leq \frac{1}{\varpi} e^{-\frac{M^2 \mathsf{gap}}{4}\left(\frac{\underline{\kappa}}{\bar{\kappa}}\right)^2 \sigma_n \epsilon^2} \sum_{\delta \in \Delta_s} \left(\frac{1}{p^{\mathsf{u}}}\sqrt{\frac{\rho_1}{2\pi}}\right)^{\|\delta\|_0} \left(2\pi\rho_1^{-1}\right)^{\|\delta\|_0/2} \\ &\leq 2e^{s_\star(\mathsf{u}+1)\log(p)} e^{-\frac{M^2 \mathsf{gap}}{4}\left(\frac{\underline{\kappa}}{\bar{\kappa}}\right)^2 \sigma_n \epsilon^2} \leq 2e^{-\frac{M^2}{8\mathsf{gap}}\left(\frac{\underline{\kappa}}{\bar{\kappa}}\right)^2 \sigma_n \mathsf{r}_1^2},\end{aligned}$$

under the sample size condition (11), where the third inequality follows from the assumptions $\mathsf{u} > 1$, and $p^{\mathsf{u}-1} > 2$. This proves the theorem.

□

We derive here a lower bound on the normalizing constant of the quasi-posterior distribution.

Lemma S-1. *Suppose that the dataset* $\mathbf{Z}$ *satisfies Assumption H3, and* $1 < \sigma_n \leq p$. *Then we can an absolute constant* c_0 *such that* $p \geq \max(c_0, e^1 s_\star)$, *we have*

$$\sum_{\delta \in \Delta_s} \left(\frac{1}{p^{\mathsf{u}}}\sqrt{\frac{\rho_1}{2\pi}}\right)^{\|\delta\|_0} \int_{\mathbb{R}^{\|\delta\|_0}} \exp\left(-\frac{\rho_1}{2}\|u\|_2^2 + \sigma_n \bar{\mathsf{R}}_n((u,0)_\delta; \mathbf{Z})\right) \mathrm{d}u \;\geq\; e^{-s_\star(\mathsf{u}+1)\log(p))} \quad \text{(S-3)}$$

Proof. Clearly, the left hand side of (S-3) is bounded from below by

$$\left(\frac{1}{p^{\mathsf{u}}}\sqrt{\frac{\rho_1}{2\pi}}\right)^{s_\star} \int_{\mathbb{R}^{s_\star}} \exp\left(-\frac{\rho_1}{2}\|u\|_2^2 + \sigma_n \bar{\mathsf{R}}_n((u,0)_{\delta_\star}; \mathbf{Z})\right) \mathrm{d}u.$$

For any $\theta \in \mathbb{R}^p$ that has the same support as $\theta_\star$, we have

$$\begin{aligned}\bar{\mathsf{R}}_n(\theta; \mathbf{Z}) &= \frac{\theta^{\mathrm{T}} \hat{A} \theta}{\theta^{\mathrm{T}} \hat{B} \theta} - \frac{\theta_\star^{\mathrm{T}} \hat{A} \theta_\star}{\theta_\star^{\mathrm{T}} \hat{B} \theta_\star} \\ &= \frac{\theta^{\mathrm{T}} \hat{\Sigma} \theta}{\theta^{\mathrm{T}} \hat{B} \theta} - \frac{\theta_\star^{\mathrm{T}} \hat{\Sigma} \theta_\star}{\theta_\star^{\mathrm{T}} \hat{B} \theta_\star} \\ &= \frac{\theta^{\mathrm{T}} \hat{\Sigma} \theta \left(\theta_\star^{\mathrm{T}} \hat{B} \theta_\star - \theta^{\mathrm{T}} \hat{B} \theta\right)}{(\theta_\star^{\mathrm{T}} \hat{B} \theta_\star)(\theta^{\mathrm{T}} \hat{B} \theta)} + \frac{1}{\theta_\star^{\mathrm{T}} \hat{B} \theta_\star} \left(\theta^{\mathrm{T}} \hat{\Sigma} \theta - \theta_\star^{\mathrm{T}} \hat{\Sigma} \theta_\star\right).\end{aligned}$$

Since $\mathsf{R}_n(\cdot;\mathbf{Z})$ is invariant to rescaling, we can assume without any loss of generality that $\|\theta\|_2 = \|\theta_\star\|_2 = 1$. Therefore for $\mathbf{Z}$ satisfying H3-(1),we have from Lemma S-3

$$\left|\bar{\mathsf{R}}_n(\theta;\mathbf{Z})\right| \;\leq\; \left(2\left(\frac{\bar\kappa}{\underline\kappa}\right)^2 + 2\left(\frac{\bar\kappa}{\underline\kappa}\right)\right)\|\theta\theta^{\mathrm{T}} - \theta_\star\theta_\star^{\mathrm{T}}\|_{\mathsf{F}} \;\leq\; 4\left(\frac{\bar\kappa}{\underline\kappa}\right)^2\|\theta\theta^{\mathrm{T}} - \theta_\star\theta_\star^{\mathrm{T}}\|_{\mathsf{F}}. \quad \text{(S-4)}$$

It follows from the above observations that for $\mathbf{Z}$ satisfying H3 the left hand side of (S-3) is bounded from below by

$$\left(\frac{1}{p^{\mathsf{u}}}\sqrt{\frac{\rho_1}{2\pi}}\right)^{s_\star}\int_{\mathbb{R}^{s_\star}}\exp\left(-\frac{\rho_1}{2}\|u\|_2^2 - \frac{C}{2}\sigma_n\|uu^{\mathrm{T}} - \theta_\star\theta_\star^{\mathrm{T}}\|_{\mathsf{F}}\right)\mathrm{d}u$$
$$\geq \left(\frac{1}{p^{\mathsf{u}}}\sqrt{\frac{\rho_1}{2\pi}}\right)^{s_\star} e^{-C\eta^2\sigma_n}\int_{\mathcal{S}_0} e^{-\frac{\rho_1}{2}\|u\|_2^2}\mathrm{d}u,$$

where $C = 8(\bar\kappa/\underline\kappa)^2$, $\eta \in (0,1)$ and $\mathcal{S}_0 \stackrel{\text{def}}{=} \{u \in \mathbb{R}^{s_\star} :\; \|uu^{\mathrm{T}} - \theta_\star\theta_\star^{\mathrm{T}}\|_{\mathsf{F}} \leq 2\eta^2\}$. Note that the integral $\int_{\mathcal{S}_0} e^{-\frac{\rho_1}{2}\|u\|_2^2}\mathrm{d}u$ is invariant to change of variables by orthogonal matrices. Hence in that integral we can replace $\theta_\star$ by the unit vector $e = (0,\ldots,0,1) \in \mathbb{R}^{s_\star}$. Using this and switching to polar coordinates, we write the integral as

$$\int_{\mathcal{S}_0} e^{-\frac{\rho_1}{2}\|u\|_2^2}\mathrm{d}u = \int_0^{+\infty} e^{-\frac{\rho_1}{2}r^2} r^{s_\star-1}\mathrm{d}r \times \nu\left(\theta \in \mathcal{S}^{s_\star-1} :\; |\sin(\theta)| \leq \eta\right),$$

where ν is the surface measure on the unit sphere $\mathcal{S}^{s_\star-1} = \{u \in \mathbb{R}^{s_\star} :\; \|u\|_2 = 1\}$, and $\sin(\theta)$ is the sine of the angle between θ and e. The measure $\nu\left(\theta \in \mathcal{S}^{s_\star-1} :\; |\sin(\theta)| \leq \eta\right)$ is equal to twice the spherical cap around the pole e defined by η. We use the formula of the spherical cap from ([10]) to write

$$\nu\left(\theta \in \mathcal{S}^{s_\star-1} :\; |\sin(\theta)| \leq \eta\right) = \frac{4\pi^{\frac{s_\star-1}{2}}}{\Gamma\left(\frac{s_\star-1}{2}\right)}\int_0^{\arcsin(\eta)}\sin^{s_\star-2}(\theta)\mathrm{d}\theta$$
$$= \frac{4\pi^{\frac{s_\star-1}{2}}}{\Gamma\left(\frac{s_\star-1}{2}\right)}\int_0^{\eta}\frac{x^{s_\star-2}}{\sqrt{1-x^2}}\mathrm{d}x \geq \frac{4\pi^{\frac{s_\star-1}{2}}}{\Gamma\left(\frac{s_\star-1}{2}\right)}\frac{\eta^{s_\star-1}}{s_\star-1}.$$

Whereas,

$$\int_0^{+\infty} e^{-\frac{\rho_1}{2}r^2} r^{s_\star-1}\mathrm{d}r = \frac{1}{2}\left(\frac{2}{\rho_1}\right)^{\frac{s_\star}{2}}\Gamma\left(\frac{s_\star}{2}\right).$$

It follows that

$$\int_{\mathcal{S}_0} e^{-\frac{\rho_1}{2}\|u\|_2^2} \mathrm{d}u \geq \frac{2}{s_\star \sqrt{\pi}} \left(\frac{2\pi}{\rho_1}\right)^{\frac{s_\star}{2}} \eta^{s_\star - 1}.$$

We conclude that for $\mathbf{Z}$ satisfying H3, and any $\eta \in (0,1)$, the left hand side of (S-3) is bounded from below by

$$\begin{aligned}\frac{2}{\sqrt{\pi} s_\star} \left(\frac{1}{p^{\mathsf{u}}}\right)^{s_\star} & e^{-(s_\star - 1)\log(1/\eta)} e^{-C\eta^2 \sigma_n} \\ &= \frac{2}{\sqrt{\pi}} \frac{p}{s_\star e^1} \exp\left(-(\mathsf{u} s_\star + 1)\log(p) - (s_\star - 1)\log(\sqrt{C\sigma_n})\right) \\ &\geq e^{-(\mathsf{u}+1)s_\star \log(p \vee C\sigma_n)},\end{aligned}$$

by taking $\eta = 1/\sqrt{C\sigma_n}$, and assuming that $p \geq e^1 s_\star$, and $\sqrt{C\sigma_n} \leq p$. This concludes the proof. □

We make use of the following version of the Davis-Kahan $\sin\Theta$ theorem taken from [15] Lemma 4.2.

Lemma S-2. *Let A be a $p \times p$ symmetric semipositive definite matrix and suppose that its eigenvalues satisfies $\lambda_1(A) > \lambda_2(A) \geq \ldots \geq \lambda_p(A)$. If a unit vector u is an eigenvector of A associated to the largest eigenvalue $\lambda_1(A)$, for all $v \in \mathbb{R}^p$, $\|v\|_2 = 1$ it holds*

$$\langle A, uu' - vv' \rangle \geq \frac{1}{2}\left(\lambda_1(A) - \lambda_2(A)\right) \|uu' - vv'\|_F^2.$$

We will need the following technical result.

Lemma S-3. *For any unit vectors u, v and square matrix B with matching dimensions, we have*

$$|\langle B, uu^{\mathrm{T}} - vv^{\mathrm{T}} \rangle| \leq 2\|B\|_{op} \|uu^{\mathrm{T}} - vv^{\mathrm{T}}\|_F, \tag{S-5}$$

Proof. Indeed, we have

$$|\langle B, uu^{\mathrm{T}} - vv^{\mathrm{T}} \rangle| = |(u-v)^{\mathrm{T}} B u + v^{\mathrm{T}} B (u-v)| \leq 2\|B\|_{\mathsf{op}} \|u - v\|_2.$$

Similarly, we have $|\langle B, uu^{\mathrm{T}} - vv^{\mathrm{T}} \rangle| \leq 2\|B\|_{\mathsf{op}} \|u + v\|_2$. Hence

$$|\langle B, uu^{\mathrm{T}} - vv^{\mathrm{T}} \rangle| \leq 2\|B\|_{\mathsf{op}} \min\left(\|u - v\|_2, \|u + v\|_2\right).$$

The result follows by noting that

$$\|uu^{\mathrm{T}} - vv^{\mathrm{T}}\|_{\mathsf{F}} \geq \min\left(\|u-v\|_2, \|u+v\|_2\right). \tag{S-6}$$

To see this, note that $\|uu^{\mathrm{T}} - vv^{\mathrm{T}}\|_{\mathsf{F}} = \|u-v\|_2\|u+v\|_2/\sqrt{2} = \|u-v\|_2\sqrt{2 - \|u-v\|_2^2/2}$. Hence, if $\|u-v\|_2^2 \leq 2$, then we have $\|uu^{\mathrm{T}} - vv^{\mathrm{T}}\|_{\mathsf{F}} \geq \|u-v\|_2$. But if $\|u-v\|_2^2 > 2$ then $\|uu^{\mathrm{T}} - vv^{\mathrm{T}}\|_{\mathsf{F}} > \|u+v\|_2$. Hence the result. □

The next result describes the behavior of the Rayleigh quotient function that yields the posterior contraction result.

Lemma S-4. *Assume H3. For any $\theta \in \mathbb{R}^p$ such that $\|\theta\|_0 \leq s$, we have*

$$R_n(\theta; \mathbf{Z}) - R_n(\theta_\star; \mathbf{Z}) \leq -\frac{\mathsf{gap}}{2}\left(\frac{\underline{\kappa}}{\bar{\kappa}}\right)^2 \|\theta\theta^{\mathrm{T}} - \theta_\star\theta_\star^{\mathrm{T}}\|_F^2 + c_0 r_1 \|\theta\theta^{\mathrm{T}} - \theta_\star\theta_\star^{\mathrm{T}}\|_F. \tag{S-7}$$

Proof. Fix $\theta \in \mathbb{R}^p$ such that $\|\theta\|_0 \leq s$. Since the Rayleigh quotient is invariant under rescaling we can assume without any loss of generality that $\|\theta\|_2 = 1$. We have

$$\begin{aligned}\bar{\mathsf{R}}_n(\theta; \mathbf{Z}) = \mathsf{R}_n(\theta; \mathbf{Z}) - \mathsf{R}_n(\theta_\star; \mathbf{Z}) &= \frac{\theta^{\mathrm{T}}\hat{\Sigma}\theta}{\theta^{\mathrm{T}}\hat{B}\theta} - \frac{\theta_\star^{\mathrm{T}}\hat{\Sigma}\theta_\star}{\theta_\star^{\mathrm{T}}\hat{B}\theta_\star} = \frac{\theta^{\mathrm{T}}\Sigma\theta}{\theta^{\mathrm{T}}B\theta} - \frac{\theta_\star^{\mathrm{T}}\Sigma\theta_\star}{\theta_\star^{\mathrm{T}}B\theta_\star} \\ &+ \left\langle \hat{\Sigma} - \Sigma, \frac{\theta\theta^{\mathrm{T}}}{\theta^{\mathrm{T}}B\theta} - \frac{\theta_\star\theta_\star^{\mathrm{T}}}{\theta_\star^{\mathrm{T}}B\theta_\star} \right\rangle + \left\langle \hat{\Sigma}, \left[\frac{\theta\theta^{\mathrm{T}}}{\theta^{\mathrm{T}}\hat{B}\theta} - \frac{\theta\theta^{\mathrm{T}}}{\theta^{\mathrm{T}}B\theta}\right] - \left[\frac{\theta_\star\theta_\star^{\mathrm{T}}}{\theta_\star^{\mathrm{T}}\hat{B}\theta_\star} - \frac{\theta_\star\theta_\star^{\mathrm{T}}}{\theta_\star^{\mathrm{T}}B\theta_\star}\right] \right\rangle. \end{aligned} \tag{S-8}$$

Set $S \stackrel{\text{def}}{=} B^{-1/2}\Sigma B^{-1/2}$, $w = B^{1/2}\theta/\|B^{1/2}\theta\|_2$, $w_\star = B^{1/2}\theta_\star/\|B^{1/2}\theta_\star\|_2$, and note that $w_\star$ is an eigenvector of S associated to the largest eigenvalue of S. Hence by the curvature lemma (Lemma S-2) we have

$$\frac{\theta^{\mathrm{T}}\Sigma\theta}{\theta^{\mathrm{T}}B\theta} - \frac{\theta_\star^{\mathrm{T}}\Sigma\theta_\star}{\theta_\star^{\mathrm{T}}B\theta_\star} = \langle S, ww^{\mathrm{T}} - w_\star w_\star^{\mathrm{T}}\rangle \leq -\frac{\mathsf{gap}}{2}\|ww^{\mathrm{T}} - w_\star w_\star^{\mathrm{T}}\|_{\mathsf{F}}^2.$$

Let $\mathsf{I} \subseteq \{1, \ldots, p\}$ be the joint support of θ and $\theta_\star$ (hence $\|\mathsf{I}\|_0 \leq s + s_\star$). Then we can express

$$\|ww^{\mathrm{T}} - w_\star w_\star^{\mathrm{T}}\|_{\mathsf{F}} = \left\| (B_{\mathsf{I},\mathsf{I}})^{1/2} \left(\frac{\theta_{\mathsf{I}}\theta_{\mathsf{I}}^{\mathrm{T}}}{\theta_{\mathsf{I}}^{\mathrm{T}}(B_{\mathsf{I},I})\theta_{\mathsf{I}}} - \frac{\theta_{\star\mathsf{I}}\theta_{\star\mathsf{I}}^{\mathrm{T}}}{\theta_{\star\mathsf{I}}^{\mathrm{T}}(B_{\mathsf{I},\mathsf{I}})\theta_{\star\mathsf{I}}} \right) (B_{\mathsf{I},\mathsf{I}})^{1/2} \right\|_{\mathsf{F}}.$$

We recall that for any square matrix A and invertible matrix B,

$$\|A\|_{\mathsf{F}} = \|B^{-1/2}B^{1/2}AB^{1/2}B^{-1/2}\|_{\mathsf{F}} \leq \|B^{-1/2}\|_{\mathsf{op}}^2 \|B^{1/2}AB^{1/2}\|_{\mathsf{F}},$$

where $\|M\|_{\mathsf{op}}$ denotes the operator norm of M. With these observations in mind, we get

$$\|ww^{\mathrm{T}} - w_\star w_\star^{\mathrm{T}}\|_{\mathsf{F}} \geq \frac{1}{\|(B_{\mathsf{I},\mathsf{I}})^{-1/2}\|_{\mathsf{op}}^2} \left\| \frac{\theta_{\mathsf{I}}\theta_{\mathsf{I}}^{\mathrm{T}}}{\theta_{\mathsf{I}}^{\mathrm{T}}(B_{\mathsf{I},I})\theta_{\mathsf{I}}} - \frac{\theta_{\star\mathsf{I}}\theta_{\star\mathsf{I}}^{\mathrm{T}}}{\theta_{\star\mathsf{I}}^{\mathrm{T}}(B_{\mathsf{I},\mathsf{I}})\theta_{\star\mathsf{I}}} \right\|_{\mathsf{F}}$$
$$\geq \underline{\kappa} \left\| \frac{\theta_{\mathsf{I}}\theta_{\mathsf{I}}^{\mathrm{T}}}{\theta_{\mathsf{I}}^{\mathrm{T}}(B_{\mathsf{I},\mathsf{I}})\theta_{\mathsf{I}}} - \frac{\theta_{\star\mathsf{I}}\theta_{\star\mathsf{I}}^{\mathrm{T}}}{\theta_{\star\mathsf{I}}^{\mathrm{T}}(B_{\mathsf{I},\mathsf{I}})\theta_{\star\mathsf{I}}} \right\|_{\mathsf{F}}.$$

We note also that for any unit vectors u, v and symmetric invertible matrix B with matching dimension,

$$\left\| \frac{uu^{\mathrm{T}}}{u^{\mathrm{T}}Bu} - \frac{vv^{\mathrm{T}}}{v^{\mathrm{T}}Bv} \right\|_{\mathsf{F}}^2 = \frac{(u^{\mathrm{T}}Bu - v^{\mathrm{T}}Bv)^2}{(u^{\mathrm{T}}Bu)^2(v^{\mathrm{T}}Bv)^2} + \frac{\|uu^{\mathrm{T}} - vv^{\mathrm{T}}\|_{\mathsf{F}}^2}{(u^{\mathrm{T}}Bu)(v^{\mathrm{T}}Bv)} \geq \frac{\|uu^{\mathrm{T}} - vv^{\mathrm{T}}\|_{\mathsf{F}}^2}{(u^{\mathrm{T}}Bu)(v^{\mathrm{T}}Bv)}. \tag{S-9}$$

Hence, under H3,

$$\|ww^{\mathrm{T}} - w_\star w_\star^{\mathrm{T}}\|_{\mathsf{F}}^2 \geq \left(\frac{\underline{\kappa}}{\bar{\kappa}}\right)^2 \|\theta_{\mathsf{I}}\theta_{\mathsf{I}}^{\mathrm{T}} - \theta_{\star\mathsf{I}}\theta_{\star\mathsf{I}}^{\mathrm{T}}\|_{\mathsf{F}}^2 = \left(\frac{\underline{\kappa}}{\bar{\kappa}}\right)^2 \|\theta\theta^{\mathrm{T}} - \theta_\star\theta_\star^{\mathrm{T}}\|_{\mathsf{F}}^2.$$

In conclusion we have

$$\frac{\theta^{\mathrm{T}}\Sigma\theta}{\theta^{\mathrm{T}}B\theta} - \frac{\theta_\star^{\mathrm{T}}\Sigma\theta_\star}{\theta_\star^{\mathrm{T}}B\theta_\star} \leq -\frac{\mathsf{gap}}{2}\left(\frac{\underline{\kappa}}{\bar{\kappa}}\right)^2 \|\theta\theta^{\mathrm{T}} - \theta_\star\theta_\star^{\mathrm{T}}\|_{\mathsf{F}}^2. \tag{S-10}$$

The second term from (S-8) can be written as

$$\left|\left\langle \hat{\Sigma} - \Sigma, \frac{\theta\theta^{\mathrm{T}}}{\theta^{\mathrm{T}}B\theta} - \frac{\theta_\star\theta_\star^{\mathrm{T}}}{\theta_\star^{\mathrm{T}}B\theta_\star} \right\rangle\right| = \left|\left\langle (\hat{\Sigma})_{\mathsf{I},\mathsf{I}} - \Sigma_{\mathsf{I},\mathsf{I}}, \frac{\frac{\theta_{\mathsf{I}}\theta_{\mathsf{I}}^{\mathrm{T}}}{\theta^{\mathrm{T}}B\theta} - \frac{\theta_{\star\mathsf{I}}\theta_{\star\mathsf{I}}^{\mathrm{T}}}{\theta_\star^{\mathrm{T}}B\theta_\star}}{\left\|\frac{\theta_{\mathsf{I}}\theta_{\mathsf{I}}^{\mathrm{T}}}{\theta^{\mathrm{T}}B\theta} - \frac{\theta_{\star\mathsf{I}}\theta_{\star\mathsf{I}}^{\mathrm{T}}}{\theta_\star^{\mathrm{T}}B\theta_\star}\right\|_{\mathsf{F}}} \right\rangle\right| \left\| \frac{\theta_{\mathsf{I}}\theta_{\mathsf{I}}^{\mathrm{T}}}{\theta^{\mathrm{T}}B\theta} - \frac{\theta_{\star\mathsf{I}}\theta_{\star\mathsf{I}}^{\mathrm{T}}}{\theta_\star^{\mathrm{T}}B\theta_\star} \right\|_{\mathsf{F}}$$
$$\leq \max_{M\in\mathbb{R}^{\mathsf{I}\times\mathsf{I}}:\, \|M\|_{\mathsf{F}}=1,\, \mathsf{Rank}(M)\leq 2} \left|\left\langle (\hat{\Sigma})_{\mathsf{I},\mathsf{I}} - \Sigma_{\mathsf{I},\mathsf{I}}, M \right\rangle\right| \times \left\| \frac{\theta_{\mathsf{I}}\theta_{\mathsf{I}}^{\mathrm{T}}}{\theta^{\mathrm{T}}B\theta} - \frac{\theta_{\star\mathsf{I}}\theta_{\star\mathsf{I}}^{\mathrm{T}}}{\theta_\star^{\mathrm{T}}B\theta_\star} \right\|_{\mathsf{F}}.$$

And we note from (S-9) and Lemma S-3 that for $\mathbf{Z}$ satisfying H3,

$$\left\| \frac{\theta_{\mathsf{I}}\theta_{\mathsf{I}}^{\mathrm{T}}}{\theta^{\mathrm{T}}B\theta} - \frac{\theta_{\star\mathsf{I}}\theta_{\star\mathsf{I}}^{\mathrm{T}}}{\theta_\star^{\mathrm{T}}B\theta_\star} \right\|_{\mathsf{F}}^2 \leq \frac{1}{\underline{\kappa}^4}\left\langle B_{\mathsf{II}}, \theta_{\mathsf{I}}\theta_{\mathsf{I}}^T - \theta_{\star,\mathsf{I}}\theta_{\star,\mathsf{I}}^{\mathrm{T}} \right\rangle^2 + \frac{1}{\underline{\kappa}^2}\left\|\theta_{\mathsf{I}}\theta_{\mathsf{I}}^T - \theta_{\star,\mathsf{I}}\theta_{\star,\mathsf{I}}^{\mathrm{T}}\right\|_{\mathsf{F}}^2$$
$$\leq \frac{1}{\underline{\kappa}^2}\left(1 + 2\left(\frac{\bar{\kappa}}{\underline{\kappa}}\right)^2\right)\|\theta\theta^T - \theta_\star\theta_\star^{\mathrm{T}}\|_{\mathsf{F}}^2. \tag{S-11}$$

Therefore for $\mathbf{Z}$ satisfying H3,

$$\left|\left\langle \hat{\Sigma}-\Sigma, \frac{\theta\theta^{\mathrm{T}}}{\theta^{\mathrm{T}}B\theta}-\frac{\theta_\star\theta_\star^{\mathrm{T}}}{\theta_\star^{\mathrm{T}}B\theta_\star}\right\rangle\right| \leq c_0\mathsf{r}_1\|\theta\theta^T-\theta_\star\theta_\star^{\mathrm{T}}\|_{\mathsf{F}}. \tag{S-12}$$

We process the last term in (S-8) as follows.

$$\begin{aligned}
&\left\langle \hat{\Sigma}, \left[\frac{\theta\theta^{\mathrm{T}}}{\theta^{\mathrm{T}}\hat{B}\theta}-\frac{\theta\theta^{\mathrm{T}}}{\theta^{\mathrm{T}}B\theta}\right]-\left[\frac{\theta_\star\theta_\star^{\mathrm{T}}}{\theta_\star^{\mathrm{T}}\hat{B}\theta_\star}-\frac{\theta_\star\theta_\star^{\mathrm{T}}}{\theta_\star^{\mathrm{T}}B\theta_\star}\right]\right\rangle \\
&\qquad = \frac{\theta^{\mathrm{T}}\hat{\Sigma}\theta}{\theta^{\mathrm{T}}\hat{B}\theta}\left\langle B-\hat{B}, \frac{\theta\theta^{\mathrm{T}}}{\theta^{\mathrm{T}}B\theta}\right\rangle - \frac{\theta_\star^{\mathrm{T}}\hat{\Sigma}\theta_\star}{\theta_\star^{\mathrm{T}}\hat{B}\theta_\star}\left\langle B-\hat{B}, \frac{\theta_\star\theta_\star^{\mathrm{T}}}{\theta_\star^{\mathrm{T}}B\theta_\star}\right\rangle \\
&= \left(\frac{\theta^{\mathrm{T}}\hat{\Sigma}\theta}{\theta^{\mathrm{T}}\hat{B}\theta}-\frac{\theta_\star^{\mathrm{T}}\hat{\Sigma}\theta_\star}{\theta_\star^{\mathrm{T}}\hat{B}\theta_\star}\right)\left\langle B-\hat{B}, \frac{\theta\theta^{\mathrm{T}}}{\theta^{\mathrm{T}}B\theta}\right\rangle + \frac{\theta_\star^{\mathrm{T}}\hat{\Sigma}\theta_\star}{\theta_\star^{\mathrm{T}}\hat{B}\theta_\star}\left\langle B-\hat{B}, \frac{\theta\theta^{\mathrm{T}}}{\theta^{\mathrm{T}}B\theta}-\frac{\theta_\star\theta_\star^{\mathrm{T}}}{\theta_\star^{\mathrm{T}}B\theta_\star}\right\rangle.
\end{aligned} \tag{S-13}$$

Hence for $\mathbf{Z}$ satisfying H3, the first term in the last display can be bounded, similar to (S-4), as

$$\begin{aligned}
&\left|\left(\frac{\theta^{\mathrm{T}}\hat{\Sigma}\theta}{\theta^{\mathrm{T}}\hat{B}\theta}-\frac{\theta_\star^{\mathrm{T}}\hat{\Sigma}\theta_\star}{\theta_\star^{\mathrm{T}}\hat{B}\theta_\star}\right)\left\langle B-\hat{B}, \frac{\theta\theta^{\mathrm{T}}}{\theta^{\mathrm{T}}B\theta}\right\rangle\right| \\
&\qquad \leq \frac{1}{\underline{\kappa}}\lambda_{\mathsf{max}}(\hat{B}-B,s)\left[4\left(\frac{\bar{\kappa}}{\underline{\kappa}}\right)^2\|\theta\theta^{\mathrm{T}}-\theta_\star\theta_\star^{\mathrm{T}}\|_{\mathsf{F}}\right] \\
&\qquad\qquad \leq \frac{4}{\underline{\kappa}}\lambda_{\mathsf{max}}(\hat{B}-B,s)\left(\frac{\bar{\kappa}}{\underline{\kappa}}\right)^2\|\theta\theta^{\mathrm{T}}-\theta_\star\theta_\star^{\mathrm{T}}\|_{\mathsf{F}}.
\end{aligned}$$

The rightmost of (S-13) is similar to (S-12):

$$\left|\frac{\theta_\star^{\mathrm{T}}\hat{\Sigma}\theta_\star}{\theta_\star^{\mathrm{T}}\hat{B}\theta_\star}\left\langle B-\hat{B}, \frac{\theta\theta^{\mathrm{T}}}{\theta^{\mathrm{T}}B\theta}-\frac{\theta_\star\theta_\star^{\mathrm{T}}}{\theta_\star^{\mathrm{T}}B\theta_\star}\right\rangle\right| \leq c_0\mathsf{r}_1\|\theta\theta^T-\theta_\star\theta_\star^{\mathrm{T}}\|_{\mathsf{F}}.$$

In conclusion the last term in (S-8) is bounded from above by

$$\left|\left\langle \hat{\Sigma}, \left[\frac{\theta\theta^{\mathrm{T}}}{\theta^{\mathrm{T}}\hat{B}\theta}-\frac{\theta\theta^{\mathrm{T}}}{\theta^{\mathrm{T}}B\theta}\right]-\left[\frac{\theta_\star\theta_\star^{\mathrm{T}}}{\theta_\star^{\mathrm{T}}\hat{B}\theta_\star}-\frac{\theta_\star\theta_\star^{\mathrm{T}}}{\theta_\star^{\mathrm{T}}B\theta_\star}\right]\right\rangle\right| \leq c_0\mathsf{r}_1\|\theta\theta^T-\theta_\star\theta_\star^{\mathrm{T}}\|_{\mathsf{F}}. \tag{S-14}$$

We conclude from (S-10-S-14) that for $\mathbf{Z}$ satisfying H3

$$\mathsf{R}_n(\theta;\mathbf{Z})-\mathsf{R}_n(\theta_\star;\mathbf{Z}) \leq -\frac{\mathsf{gap}}{2}\left(\frac{\underline{\kappa}}{\bar{\kappa}}\right)^2\|\theta\theta^{\mathrm{T}}-\theta_\star\theta_\star^{\mathrm{T}}\|_{\mathsf{F}}^2 + c_0\mathsf{r}_1\|\theta\theta^T-\theta_\star\theta_\star^{\mathrm{T}}\|_{\mathsf{F}}.$$

This ends the proof. □

S-2 Proof of Proposition 3

We present the details of this claim for $\hat{\Sigma}$, the argument being similar for the other two covariance matrices. For any $J \subset [1:p]$ of size s, we have

$$\|\hat{\Sigma}_{J,J} - \Sigma_{J,J}\|_{\mathsf{op}} = \|\Sigma_{J,J}^{1/2}\left(\Sigma_{J,J}^{-1/2}\hat{\Sigma}_{J,J}\Sigma_{J,J}^{-1/2} - I_s\right)\Sigma_{J,J}^{1/2}\|_{\mathsf{op}} \leq \|\Sigma_{J,J}^{1/2}\|_{\mathsf{op}}^2 \times \|\frac{1}{n}\sum_{i=1}^{n} U_{iJ}U_{iJ}^{\mathrm{T}} - I_s\|_{\mathsf{op}},$$

where $U_{iJ} \stackrel{\mathrm{def}}{=} \Sigma_{J,J}^{-1/2} Z_{iJ}$, where $Z_{iJ} = (Z_{ij})_{j\in J}$, is mean zero and isotropic. By Theorem 4.6.1 (Equation 4.22) of ([14]), provided that $n \geq 4c_0 s\log(p)$ for some absolute constant $c_0 > 1$, we have

$$\|\hat{\Sigma}_{J,J} - \Sigma_{J,J}\|_{\mathsf{op}} \leq CK^2\|\Sigma_{J,J}^{1/2}\|_{\mathsf{op}}^2\sqrt{\frac{c_0 s\log(p)}{n}},$$

with probability at least $1 - 2p^{-c_0 s}$. Therefore, for any matrix $A \in \mathbb{R}^{s\times s}$, with $\|A\|_{\mathsf{F}} = 1$, and $\mathsf{Rank}(A) \leq \alpha$, using the singular value decomposition of A, we have

$$\max_{\substack{A\in\mathbb{R}^{s\times s}:\ \|A\|_{\mathsf{F}}=1 \\ \mathsf{Rank}(A)\leq\alpha}} \left|\left\langle \hat{\Sigma}_{J,J} - \Sigma_{J,J}, A\right\rangle\right| \leq \sqrt{\alpha}\|\hat{\Sigma}_{J,J} - \Sigma_{J,J}\|_{\mathsf{op}} \leq CK^2\lambda_{\mathsf{max}}(\Sigma, s)\sqrt{\frac{c_0\alpha s\log(p)}{n}}.$$

Since the number of subsets of $[1:p]$ of size s is smaller than p^s, we conclude with a union bound argument that

$$\lambda_{\mathsf{max}}^{(\alpha)}(\hat{\Sigma} - \Sigma, s) \leq CK^2\lambda_{\mathsf{max}}(\Sigma, s)\sqrt{\frac{c_0\alpha s\log(p)}{n}},$$

with probability $1 - 2p^{-(c_0-1)s}$.

S-3 MCMC sampling

We sample from the simulated tempering distribution (15) using a Metropolis-within-Gibbs strategy. We describe here one iteration of the algorithm, and its transition kernel. Given (δ, θ, k), we perform a three-step update. First, given k and δ, we update θ. We let $[\theta]_\delta$ to denote the δ-selected component of θ listed in their original order: $[\theta]_\delta \stackrel{\mathrm{def}}{=} (\theta_j : j \in \{1 \leq k \leq p : \delta_k = 1\})$, and $[\theta]_{\delta^c} \stackrel{\mathrm{def}}{=} (\theta_j : j \in \{1 \leq k \leq p : \delta_k = 0\})$. We employ the fact that the selected components $[\theta]_\delta$ and the un-selected components $[\theta]_{\delta^c}$ of θ are independent

Algorithm 1 Simulated tempering for sparse canonical correlation analysis

Model Input: Matrices $\hat{A}, \hat{B}$, prior parameters $\rho_0, \rho_1, \mathsf{u}$.
MCMC Input: Number of iterations N, batch size J, temperatures $1 = t_1 < \ldots < t_K$, weights $\{c_1, \ldots, c_K\}$, and step-sizes $\{\eta_1, \ldots, \eta_K\}$.
Initialization: Set the temperature index $k^{(0)} = 1$. Draw $\delta_j^{(0)} \overset{i.i.d.}{\sim} \mathbf{Ber}(0.5)$ for $j = 1, \ldots, p$, and independently draw $\theta^{(0)} \sim \mathbf{N}(0, I_p)$.
for $t = 0$ to $N - 1$, given $(\delta^{(t)}, \theta^{(t)}, k^{(t)}) = (\delta, \theta, k)$ **do**

1. **Update** θ: Draw the components of $[\bar{\theta}]_{\bar{\delta}^c}$independently from $\mathbf{N}(0, \rho_0^{-1} t_k)$. Draw $[\bar{\theta}]_{\bar{\delta}} \sim P_{k,\delta}([\theta]_\delta, \cdot)$, where $P_{k,\delta}$ denotes the transition kernel of the MALA with step-size η_k and invariant distribution given by (S-15).

2. **Update** δ: Uniformly randomly select a subset J from $\{1, \ldots, p\}$ of size J without replacement, and draw $\bar{\delta} \sim Q_{k,\bar{\theta}}^{(\mathsf{J})}(\delta, \cdot)$, where the transition kernel described in (S-18).

3. **Update** k: Draw $\bar{k} \sim T_{\bar{\delta},\bar{\theta}}(k, \cdot)$, where $T_{\delta,\theta}$ is the transition kernel of the Metropolis-Hastings on $\{1, \ldots, K\}$ with invariant distribution given by (S-19) and random walk proposal that has reflection at the boundaries.

4. **New MCMC state**: Set $(\delta^{(t+1)}, \theta^{(t+1)}, k^{(t+1)}) = (\bar{\delta}, \bar{\theta}, \bar{k})$.

end for
Output: $\{(\delta^{(t)}, \theta^{(t)}, k^{(t)}) : \ 0 \leq t \leq N \quad \text{s.t.} \quad k^{(t)} = 1\}$

conditional on k and δ to update θ. In addition, given k and δ, the components of $[\theta]_{\delta^c}$ are i.i.d. $\mathbf{N}(0, t_k \rho_0^{-1})$ and the distribution of $[\theta]_\delta$ has density on $\mathbb{R}^{\|\delta\|_0}$ proportional to

$$u \mapsto \exp\left(-\frac{\rho_1}{2t_k}\|u\|_2^2 + \frac{\sigma_n}{t_k}\mathsf{R}_n((u,0)_\delta; \mathbf{Z})\right), \tag{S-15}$$

where the notation $(u, 0)_\delta$ denotes the vector in $\mathbb{R}^p$ such that $[(u, 0)_\delta]_\delta = u$. Hence we update $[\theta]_\delta$ using a Metropolis adjusted Langevin algorithm (MALA) on $\mathbb{R}^{\|\delta\|_0}$ with target distribution (S-15), and step-size η_k (we use different step-sizes for different temperature levels). Let $M_{\delta,k}$ denote the resulting transition kernel on $\mathbb{R}^{\|\delta\|_0}$. For more details on the MALA, see e.g., [13]. For convenience, we write $P_{k,\delta}$ to denote the Markov kernel on $\mathbb{R}^p$

corresponding to the update of θ just described. Specifically,

$$P_{\delta,k}(\theta, \mathrm{d}\theta') = M_{\delta,k}([\theta]_\delta, \mathrm{d}[\theta']_\delta) \prod_{j:\ \delta_j=0} \mathbf{N}(0, \rho_0^{-1})(\mathrm{d}\theta'_j),$$

where $\mathbf{N}(\mu, \sigma^2)(\mathrm{d}x)$ is a short for the Gaussian measure on $\mathbb{R}$ with mean μ and variance σ^2.

Secondly, we update δ by applying a Gibbs sampler to the conditional distribution of δ given k and θ. Note that the conditional distribution of δ_j given k, θ and δ_{-j}, where $\delta_{-j} \stackrel{\text{def}}{=} (\delta_1, \ldots, \delta_{j-1}, \delta_{j+1}, \ldots, \delta_p)$, is the Bernoulli distribution $\mathbf{Ber}(q_j)$, with probability of success given by

$$q_j \stackrel{\text{def}}{=} \left\{1 + \exp\left(-\frac{\mathsf{a}}{t_k} + \frac{1}{2t_k}(\rho_1 - \rho_0)\theta_j^2\right) \exp\left(\frac{\sigma_n}{t_k}\mathsf{R}_n(\theta_{\delta^{(j,0)}}; \mathbf{Z}) - \frac{\sigma_n}{t_k}\mathsf{R}_n(\theta_{\delta^{(j,1)}}; \mathbf{Z})\right)\right\}^{-1}, \tag{S-16}$$

where

$$\delta_i^{(j,0)} \stackrel{\text{def}}{=} \begin{cases} 0 & i = j \\ \delta_i & i \neq j \end{cases}, \quad \delta_i^{(j,1)} \stackrel{\text{def}}{=} \begin{cases} 1 & i = j \\ \delta_i & i \neq j \end{cases}. \tag{S-17}$$

Given k, θ and j, let $Q_{k,\theta}^{(j)}$ denote the transition kernel on Δ which, given δ, leaves δ_i unchanged for all $i \neq j$, and draws $\delta_j \sim \mathbf{Ber}(q_j)$. We update δ as follows: randomly draw a subset $\mathsf{J} = \{\mathsf{J}_1, \ldots, \mathsf{J}_J\}$ of size J from $\{1, \ldots, p\}$, and update δ using the transition kernel on Δ given by

$$Q_{k,\theta}^{(\mathsf{J})} \stackrel{\text{def}}{=} Q_{k,\theta}^{(\mathsf{J}_1)} Q_{k,\theta}^{(\mathsf{J}_2)} \cdots Q_{k,\theta}^{(\mathsf{J}_J)}. \tag{S-18}$$

The resulting overall kernel on Δ is

$$\bar{Q}_{k,\theta} = \sum_{\mathsf{J}:\ |\mathsf{J}|=J} \binom{p}{J}^{-1} Q_{k,\theta}^{(\mathsf{J})}.$$

Thirdly, given δ and θ, we update k using a standard Metropolis-Hastings algorithm with a random walk proposal that has reflection at the boundaries. Specifically, at k we propose with equal probability either $k-1$ or $k+1$, except at 1, where we only propose 2, and at K, where we only propose $K-1$. We write $T_{\delta,\theta}$ to denote the transition kernel on $\{1, \ldots, K\}$ of this Metropolis-Hastings algorithm with invariant distribution

$$i \mapsto \frac{1}{c_i} \exp\left\{\frac{\mathsf{a}}{t_i}\|\delta\|_0 - \frac{\rho_1}{2t_i}\|\theta_\delta\|_2^2 - \frac{\rho_0}{2t_i}\|\theta - \theta_\delta\|_2^2 + \frac{\sigma_n}{t_i}\mathsf{R}_n(\theta_\delta; \mathbf{Z})\right\}. \tag{S-19}$$

Lastly, we collect samples by retaining the values of (δ, θ) at iterations at which $k = 1$. In stationarity these samples have distribution (8).

S-3.1 Parameter choices and adaptive tuning

Throughout the simulation, we specify the parameters of the prior distribution in the following way. We let $\rho_1 = \frac{1}{2}$, and $\rho_0 = n/10$, where n is the sample size, and we set $\mathsf{u} = 1.5$.

Algorithm 1 also depends on the user-defined parameters J, K, $(t_1, \ldots, t_K)$, $(c_1, \ldots, c_K)$, and $(\eta_1, \ldots, \eta_K)$. The parameter J (the Gibbs sampling batch size) does not greatly impact performance, and setting $J = 100$ works well in most settings. Efficient selection and tuning of temperatures in simulated tempering has received some attention ([8, 2]), and despite some progress ([12]), to the best of our knowledge, there is no practical and scalable algorithm to do so. In our implementation we use variations of the geometric scaling. We refer the reader to Section 4 for specific choices.

We tune the step-sizes $\eta = (\eta_1, \ldots, \eta_K)$ of MALA and the weights $(c_1, \ldots, c_K)$ of simulated tempering using adaptive MCMC methods , see e.g., [1]. To tune η_k, we follow the algorithm proposed in [3], with a targeted acceptance probability of 30%. For simulated tempering to visit all temperature levels frequently, the weights $(c_1, \ldots, c_K)$ need to be adequately tuned. We refer the reader to [8] for an extensive discussion of the issue. This problem can be efficiently solved using the Wang-Landau algorithm for simulated tempering as developed in [4]. We follow this approach here. The fully adaptive MCMC sampler is presented in Algorithm 2.

S-4 Coupled Markov chains for mixing time estimation

At least empirically, simulated tempering is well-known to improve mixing when dealing with multimodal distributions ([8, 11]). However, rigorous results are far less well-established. Using a Markov kernel decomposition approach, ([16]) gives a lower bound on the spectral gap of simulated tempering in terms of the spectral gaps of the component kernels and the so-called projection kernel. However, applying their result to a specific problem remains non-trivial. Furthermore, their lower bound decays exponentially fast in the number of components in the partition, which clearly limits its relevance in our setting. Using a similar Markov kernel decomposition technique, ([7]) has a more explicit upper bound on the mixing time of simulated tempering. However their result applies to a different algorithm than the one considered here, and they consider a specific form of the target distribution that does not include (15).

Algorithm 2 Adaptive version of simulated tempering for Canonical correlation analysis

Model Input: Matrices $\hat{A}, \hat{B}$, prior parameters $\rho_0, \rho_1, \mathsf{u}$.
MCMC Input: Number of iterations N, Batch size J, temperatures $1 = t_1 < \ldots, < t_K$.
Adaptive MCMC Input: $a = 10$ and $w \in (0, 1)$.
MCMC Initialization: Set $k^{(0)} = 1$. Draw $\delta_j^{(0)} \overset{i.i.d.}{\sim} \mathbf{Ber}(0.5), \forall j = 1, \ldots, p$, and independently $\theta^{(0)} \sim \mathbf{N}(0, I_p)$.
Adaptation Parameters Initialization : . Set $\ell^{(0)} = \mathbf{0} \in \mathbb{R}^K$, $v^{(0)} = (0, \ldots, 0) \in \mathbb{R}^K$, $\nu^{(0)} = (0, \ldots, 0) \in \mathbb{R}^K$, and choose $c^{(0)} \in (0, \infty)^K$.
for $t = 1$ to $N - 1$, given $(\delta^{(t)}, \theta^{(t)}, k^{(t)}) = (\delta, \theta, k)$, $\ell^{(t)} = \ell$, $c^{(t)} = c$, $v^{(t)} = v$, and $\nu^{(t)} = \nu$ **do**

1. **Update θ and ℓ**: Draw the components of $[\bar{\theta}]_{\bar{\delta}^c}$independently from $\mathbf{N}(0, \rho_0^{-1} t_k)$. Draw $[\bar{\theta}]_\delta \sim P_{k,\delta}([\theta]_\delta, \cdot)$, where $P_{k,\delta}$ denotes the transition kernel of the MALA with step-size e^{ℓ_k} and invariant distribution given by (S-15). Denote α as the acceptance probability of the MALA update. Set

$$\bar{\ell}_k = \ell_k + v_k^{-0.6}(\alpha - 0.3) \quad \text{and for } i \neq k, \text{ set } \bar{\ell}_i = \ell_i.$$

2. **Update δ**: Uniformly randomly select a subset J from $\{1, \ldots, p\}$ of size J without replacement, and draw $\bar{\delta} \sim Q_{k,\bar{\theta}}^{(\mathsf{J})}(\delta, \cdot)$, where the transition kernel described in (S-18).

3. **Update k, c, v and ν**: Draw $\bar{k} \sim T_{\bar{\delta},\bar{\theta}}(k, \cdot)$, where $T_{\delta,\theta}$ is the transition kernel of the Metropolis-Hastings on $\{1, \ldots, K\}$ with invariant distribution given by (S-19) and random walk proposal with reflection at the boundaries. We then set

$$\bar{c}_{\bar{k}} = c_{\bar{k}} e^a, \quad \bar{v}_{\bar{k}} = v_{\bar{k}} + 1, \ \bar{\nu}_{\bar{k}} = \nu_{\bar{k}} + 1, \quad \text{and for } \ i \neq \bar{k}, \bar{c}_i = c_i, \ \bar{v}_i = v_i, \quad \text{and } \ \bar{\nu}_i = \nu_i.$$

4. **Update a and ν**: If $\|\nu/(\sum_{k=1}^K \nu_k) - 1/K\|_\infty \leq w/K$, then set $a = a/2, \nu = \mathbf{0} \in \mathbb{R}^K$.
5. **New MCMC state**: Set $(\delta^{(t+1)}, \theta^{(t+1)}, k^{(t+1)}) = (\bar{\delta}, \bar{\theta}, \bar{k})$, $\ell^{(t+1)} = \bar{\ell}$, $c^{(t+1)} = \bar{c}$, $v^{(t+1)} = \bar{v}$, and $\nu^{(t+1)} = \bar{\nu}$.

end for
Output: $\{(\delta^{(t)}, \theta^{(t)}, k^{(t)}) : \ 0 \leq t \leq N \quad \text{s.t.} \quad k^{(t)} = 1\}$

Given the lack of theoretical mixing time analysis of simulated tempering, we take a more empirical approach based on the unbiased Markov Chain Monte Carlo framework in ([5, 9]). Let $\{X^{(t)},\ t \geq 0\}$ be the Markov chain generated by our simulated tempering algorithm, where $X^{(t)} = (\delta^{(t)}, \theta^{(t)}, k^{(t)}) \in \mathsf{X}$. Let P denote its transition kernel (which is described in Section S-4 in the supplementary material). Following [9], we construct a coupling $\check{P}$ of P with itself: that is, a transition kernel on $\mathsf{X} \times \mathsf{X}$ such that $\check{P}((x,y), A \times \mathsf{X}) = P(x, A)$, $\check{P}((x,y), \mathsf{X} \times B) = P(y, B)$, for all $x, y \in \mathsf{X}$, and all measurable sets A, B. Furthermore, $\check{P}((x,x), \mathcal{D}) = 1$ where $\mathcal{D} \stackrel{\text{def}}{=} \{(x,x) :\ x \in \mathsf{X}\}$. The construction of the Markov kernel $\check{P}$ is described in Section S-4 in the supplementary material.

Given $\check{P}$, a lag $L \geq 1$, and an initial distribution as given in the initialization step of Algorithm 1, we simulate a bivariate Markov chain $\{(X_t, Y_{t-L}),\ t \geq L\}$ as follows. First draw $X^{(0)} \sim \Pi^{(0)}$ and $Y^{(0)} \sim \Pi^{(0)}$. Next, for $1 \leq t \leq L$, we draw $X_t | (X_0, Y_0, X_1, X_2, \ldots, X_{t-1}) \sim P(X_{t-1}, \cdot)$. Then for $t > L$, we draw

$$(X_t, Y_{t-L}) | \{(X_{t-1}, Y_{t-L-1}), \ldots, (X_L, Y_0), X_{L-1}, \ldots, X_0\} \sim \check{P}\left((X_{t-1}, Y_{t-L-1}), \cdot\right).$$

In other words, at each time $t > L$ we attempt to couple the two chains while maintaining the correct marginals. We define $\tau^{(L)} \stackrel{\text{def}}{=} \inf\{t \geq L :\ X_t = Y_{t-L}\}$, and have the following:

Proposition 5. *Let $\{X^{(t)},\ t \geq 0\}$ be the Markov chain generated by the simulated tempering algorithm, and let $\bar{\Pi}^{(t)}$ denote the distribution of $X^{(t)}$. For all $t \geq 0$, we have*

$$\|\bar{\Pi}^{(t)} - \bar{\Pi}\|_{\text{tv}} \leq \mathbb{E}\left[\max\left(0, \left\lceil \frac{\tau^{(L)} - L - t}{L} \right\rceil\right)\right]. \tag{S-20}$$

Proof. See Section S-4.2. □

This inequality implies that by simulating multiple copies of the bivariate chain, and approximating the expectation in (S-20) by Monte Carlo, we can actually estimate the mixing time of our algorithm. This gives us the possibility to investigate empirically the mixing time of our sampler with some theoretical guarantees.

S-4.1 Coupled Markov Chains

We describe here the specific coupled Markov chain employed to estimate the mixing time plots presented in Section S-4.3.2. We refer the reader to [5] and [9] for more details on the construction of such coupled kernels. We modify Algorithm 1 to construct the coupled kernel $\check{P}$. It suffices here to describe one iteration of the coupled chain. At some iteration

$t \geq 1$, suppose that $(\delta^{(1,L+t)}, \theta^{(1,L+t)}, k^{(1,L+t)}) = (\delta^{(1)}, \theta^{(1)}, k^{(1)})$ and $(\delta^{(2,t)}, \theta^{(2,t)}, k^{(2,t)}) = (\delta^{(2)}, \theta^{(2)}, k^{(2)})$.

In step 1, to update $\theta^{(1)}$ and $\theta^{(2)}$, we partition the indices $\{1, \ldots, p\}$ into four groups: $G_{ab} = \{j : \delta_j^{(1)} = a, \delta_j^{(2)} = b\}$ for $a, b = 0, 1$. To update the components of $\theta_{G_{00}}^{(1)}$ and $\theta_{G_{00}}^{(2)}$, for any $j \in G_{00}$ we first draw a common standard normal random variables Z_j, and then obtain $\bar{\theta}_j^{(i)} = t_{k^{(i)}} \rho_0^{-1} Z_j$ for $i = 1, 2$. To update the components of $\theta_{G_{01}}^{(1)}$ and $\theta_{G_{01}}^{(2)}$, for any $j \in G_{01}$ we again first draw a common standard normal random variables Z_j, and then obtain $\bar{\theta}_j^{(1)} = t_{k^{(1)}} \rho_0^{-1} Z_j$, and simultaneously draw $\bar{\theta}_j^{(2)}$ using MALA with proposal $\theta_j^{(2)} + \eta_{k^{(2)}} \nabla \log \pi(\theta_j^{(2)}) + \sqrt{2\eta_{k^{(2)}}} Z_j$, where $\pi(\theta_j^{(2)})$ is the marginal posterior distribution of $\theta_j^{(2)}$. Notice that the joint distribution of $[\theta^{(2)}]_{\delta^{(2)}}$ is given by $W_{k^{(2)}, \delta^{(2)}}$, whose density is proportional to (S-15). A similar update procedure is used for updating the components of $\theta_{G_{10}}^{(1)}$ and $\theta_{G_{10}}^{(2)}$. To update the components of $\theta_{G_{11}}^{(1)}$ and $\theta_{G_{11}}^{(2)}$, we draw reflection-coupled MALA proposals in [5], and then for the acceptation step, $\theta_{G_{11}}^{(1)}$ and $\theta_{G_{11}}^{(2)}$ share the same uniform random variables.

In step 2, to update $\delta^{(1)}$ and $\delta^{(2)}$, we first make use of the same randomly drawn subset J. For $i = 1, 2$, drawing $\bar{\delta}^{(i)} \sim Q_{k,\theta}^{(\mathsf{J})}(\delta^{(i)}, \cdot)$ is equivalent to let $\bar{\delta}_{-\mathsf{J}}^{(i)} = \delta_{-\mathsf{J}}^{(i)}$, and for any $j \in \mathsf{J}$, draw $\bar{\delta}_j^{(i)} \sim \mathbf{Ber}(q_j^{(i)})$ which we implement in the following way. We first draw a common uniform number $u_j \sim \mathbf{Uniform}(0, 1)$, then we obtain $\bar{\delta}_j^{(i)} = \mathbf{1}\{q_j^{(i)} \leq u_j\}$ for $i = 1, 2$.

In step 3, to update $k^{(1)}$ and $k^{(2)}$, we use a common uniform random number to make the proposal move, and a common uniform random number for the acceptation step.

Remark 6. Note that although the empirical mixing time estimation method of [5] described above only applies to Markov chains with fixed parameters, we have applied it here to Algorithm 2, which is an MCMC sampler with adaptively tuned parameters. We conjecture that the unbiased MCMC methodology remains approximately valid for well-constructed adaptive MCMC samplers. However the question deserves more research.

S-4.2 Proof of Proposition 5

Using the notations established in Section S-3 the transition kernel of the simulated tempering algorithm on $\mathcal{X} \stackrel{\text{def}}{=} \Delta \times \mathbb{R}^p \times \{1, \ldots, K\}$ is given by

$$P((\delta, \theta, k); (\delta', \mathrm{d}\theta', k')) = P_{\delta,k}(\theta, \mathrm{d}\theta') \left\{ \sum_{\mathsf{J}:\, |\mathsf{J}|=J} \binom{p}{J}^{-1} Q_{k,\theta'}^{(\mathsf{J})}(\delta, \delta') \right\} T_{\delta',\theta'}(k, k'),$$

and we call $\bar{P}$ the transition kernel of the coupled chain on $\mathcal{X} \times \mathcal{X}$ as described in Section S-4. The kernel P is a standard Metropolis-within-Gibbs kernel to sample from the density (15) that is positive everywhere. Therefore, P is phi-irreducible, aperiodic and has invariant distribution $\bar{\Pi}(\cdot|\mathbf{Z})$ by construction. Furthermore, for any nonempty compact set $\mathcal{C}$ of $\mathbb{R}^p$, the set $\bar{\mathcal{C}} \stackrel{\text{def}}{=} \mathcal{C} \times \Delta \times \{1, \ldots, K\}$ is a small set for P, and it is easy to see from the construction of the coupled chain that,

$$\min_{x,y \in \bar{\mathcal{C}}} \bar{P}^{n_0}\left((x,y); \mathcal{D}\right) > 0, \quad \text{with } n_0 = \max\left(K, \frac{p}{J}\right).$$

Therefore, according to Proposition 4 of ([9]) to establish the finiteness of the average meeting time $\mathbb{E}(\tau^{(L)})$, it suffices to show that there exist a drift function $V: \mathcal{X} \to [1, \infty)$, $\lambda \in (0,1)$, and $b < \infty$ such that

$$PV(x) \leq \lambda V(x) + b\mathbf{1}_{\bar{\mathcal{C}}}(x), \quad \text{for all} \quad x = (\delta, \theta, k) \in \mathcal{X}, \tag{S-21}$$

for some small $\bar{\mathcal{C}}$ of the form $\bar{\mathcal{C}}_L \stackrel{\text{def}}{=} \{x \in \mathcal{X}: \ V(x) \leq L\}$. We show (S-21) in three steps, with

$$V(\delta, \theta, k) \stackrel{\text{def}}{=} 1 + \frac{1}{t_k}\|\theta_\delta\|_2^2.$$

Step 1: Action of the kernel $T_{\delta,\theta}$ We first show that for all $(\theta, \delta, k) \in \mathcal{X}$,

$$T_{\delta,\theta}V(\delta, \theta, k) \leq V(\delta, \theta, k) + c_0, \tag{S-22}$$

for some constant c_0. To show this, we find it easy to reason in a slightly more general terms. Consider a discrete distribution on $\{1, \ldots, K\}$, given by

$$\pi(k) \propto \frac{1}{c_k} e^{-U/t_k},$$

for some increasing sequence $\{c_k, \ 1 \leq k \leq K\}$, and for some nonnegative constant U. Consider a Metropolis-Hastings algorithm to sample from π with a proposal q on $\{1, \ldots, K\}$ such that at j, we propose to move only to $j-1$ or $j+1$ for equal probability (at 1 we propose to move only to 2, and at K we propose to move only to $K-1$). Call M the transition kernel of that Metropolis-Hastings, and for some nonnegative constant B, define

$$V(j) = \frac{B}{t_j}, \ \ j = 1, \ldots, K.$$

By the definition of the Metropolis-Hasting kernel, we have

$$MV(j) = V(j) + \sum_{j'=1}^{K} (V(j') - V(j)) \min\left(1, \frac{\pi(j')q(j',j)}{\pi(j)q(j,j')}\right) q(j,j') \\ = V(j) + \sum_{j'=1}^{K} R(j,j')q(j,j'),$$

where

$$R(j,j') = (V(j') - V(j)) \min\left(1, \frac{\pi(j')q(j',j)}{\pi(j)q(j,j')}\right).$$

Note that $V(j+1) \leq V(j)$, and therefore $R(j, j+1) \leq 0$. Whereas

$$R(j,j-1) = B\left(\frac{1}{t_{j-1}} - \frac{1}{t_j}\right) \min\left(1, \frac{c_j}{c_{j-1}} \frac{q(j-1,j)}{q(j,j-1)} e^{-\left(\frac{1}{t_{j-1}} - \frac{1}{t_j}\right)U}\right) \\ \leq \left(\frac{B}{U}\right) \frac{2c_j}{c_{j-1}} \left(\frac{1}{t_{j-1}} - \frac{1}{t_j}\right) U e^{-\left(\frac{1}{t_{j-1}} - \frac{1}{t_j}\right)U} \\ \leq 2e^{-1} \left(\max_j \frac{c_j}{c_{j-1}}\right) \frac{B}{U} = C,$$

where we use the fact that $q(j',j)/q(j,j') \leq 2$, and the observation that $te^{-t} \leq e^{-1}$ for all $t \geq 0$. Using these we have

$$MV(1) = V(1) + R(1,2) \leq V(1).$$

$$MV(K) = V(K) + R(K, K-1) \leq V(K) + C.$$

For $2 \leq j \leq K-1$,

$$MV(j) = V(j) + \frac{1}{2}R(j,j-1) + \frac{1}{2}R(j,j+1) \leq V(j) + \frac{C}{2}.$$

Hence, for all $1 \leq j \leq K$, it holds

$$MV(j) \leq V(j) + 2e^{-1} \left(\max_j \frac{c_j}{c_{j-1}}\right) \frac{B}{U}.$$

We can apply this result to the kernel $T_{\delta,\theta}$ with $U = \frac{\rho_1}{2}\|\theta_\delta\|_2^2 + \frac{\rho_0}{2}\|\theta - \theta_\delta\|_2^2 - \sigma_n \mathsf{R}_n(\theta_\delta; \mathbf{Z}) - \mathsf{a}\|\delta\|_0$, and $B = \|\theta_\delta\|_2^2$. Under assumption H2, $\mathsf{R}_n(\theta_\delta; \mathbf{Z}) \leq 1$. Hence for $\rho_1\|\theta_\delta\|_2^2 \geq 4(\sigma_n + \mathsf{a}p)$, the chosen U is non-negative and we get

$$\begin{aligned} T_{\delta,\theta}V(\delta,\theta,k) &\leq V(\delta,\theta,k) \\ &\quad + 2e^{-1}\left(\max_j \frac{c_j}{c_{j-1}}\right) \frac{\|\theta_\delta\|_2^2}{\frac{\rho_1}{2}\|\theta_\delta\|_2^2 + \frac{\rho_0}{2}\|\theta - \theta_\delta\|_2^2 - \sigma_n \mathsf{R}_n(\theta_\delta; \mathbf{Z}) - \mathsf{a}\|\delta\|_0} \\ &\leq V(\delta,\theta,k) + \frac{8e^{-1}}{\rho_1}\left(\max_j \frac{c_j}{c_{j-1}}\right), \end{aligned}$$

for $\|\theta_\delta\|_2 \geq L$, for L taken large enough.

Step 2: Accounting for the kernel $\bar{Q}$ For consistency in the notation, we write summations as integrals with respect to the counting measure. Using (S-22) and the definition of $\bar{Q}$, we have for all $(\delta,\theta,k) \in \mathcal{X}$ such that $\|\theta_\delta\|_2 \geq L$ for some appropriately large L,

$$\begin{aligned} \int_\Delta \bar{Q}_{\theta,k}(\delta, \mathrm{d}\delta') \int T_{\theta,\delta'}(k, \mathrm{d}k') V(\theta,\delta',k') &\leq \int_\Delta \bar{Q}_{\theta,k}(\delta, \mathrm{d}\delta') V(\theta,\delta',k) + c_0 \\ &= \sum_{\mathsf{J}:\,|\mathsf{J}|=J} \binom{p}{J}^{-1} \int_\Delta Q^{(\mathsf{J})}_{\theta,k}(\delta, \mathrm{d}\delta') V(\theta,\delta',k) + c_0. \end{aligned}$$

Given a selection $\mathsf{J} = \{j_1, \ldots, j_J\} \subseteq \{1, \ldots, p\}$, and $j_i \in \mathsf{J}$, we have

$$\int_\Delta \tilde{Q}^{(j_i)}_{k,\theta}(\delta, \mathrm{d}\delta') V(\theta,\delta',k) \;=\; V(\theta,\delta,k) \;+\; (q_{j_i} - \delta_{j_i})\,\frac{\theta_{j_i}^2}{t_k} \;\leq\; V(\theta,\delta,k) \;+\; (1 - \delta_{j_i})\,\frac{\theta_{j_i}^2}{t_k},$$

where q_j is as given in (S-16). It follows that

$$\int_\Delta Q^{(\mathsf{J})}_{\theta,k}(\delta, \mathrm{d}\delta') V(\theta,\delta',k) \leq V(\theta,\delta,k) + \frac{1}{t_k}\sum_{i=1}^{J} \theta_{j_i}^2 \mathbf{1}_{\{\delta_{j_i}=0\}}.$$

We conclude that for $\|\theta_\delta\|_2 \geq L$,

$$\begin{aligned} \int_\Delta \bar{Q}_{\theta,k}(\delta, \mathrm{d}\delta') \int T_{\theta,\delta'}(k, \mathrm{d}k') V(\theta,\delta',k') &\leq V(\theta,\delta,k) \\ &\quad + \frac{1}{t_k} \sum_{\mathsf{J}:\,|\mathsf{J}|=J} \binom{p}{J}^{-1} \left\{ \sum_{i=1}^{J} \theta_{\mathsf{J}_i}^2 \mathbf{1}_{\{\delta_{\mathsf{J}_i}=0\}} \right\} + c_0. \quad \text{(S-23)} \end{aligned}$$

Step 3: Drift condition for P We recall that under the kernel $P_{\delta,k}$ the components $\{\theta'_j,\ j:\ \delta_j = 0\}$ are drawn independently from the Gaussian distribution $\mathbf{N}(0, \rho_0^{-1})$. Therefore, for $\|\theta_\delta\|_2 \geq L$, using (S-23), we have

$$\begin{aligned}
&\int_{\mathcal{X}} P\left((\theta,\delta,k);(\mathrm{d}\theta',\mathrm{d}\delta',\mathrm{d}k')\right) V(\theta',\delta',k') \\
&\qquad \leq \int_{\mathbb{R}^p} P_{\delta,k}(\theta,\mathrm{d}\theta') \left[V(\theta',\delta,k) + \frac{1}{t_k} \sum_{\mathsf{J}:\ |\mathsf{J}|=J} \binom{p}{J}^{-1} \left\{ \sum_{i=1}^{J} (\theta'_{\mathsf{J}_i})^2 \mathbf{1}_{\{\delta_{\mathsf{J}_i}=0\}} \right\} + c_0 \right] \\
&\qquad\qquad \leq 1 + \frac{1}{t_k} \int_{\mathbb{R}^{\|\delta\|_0}} M_{k,\delta}(\theta_\delta,\mathrm{d}v)\|v\|_2^2 + \frac{J}{\rho_0 t_k} + c_0, \quad \text{(S-24)}
\end{aligned}$$

where $M_{k,\delta}$ is the kernel of the MALA with target distribution proportional to

$$u \mapsto \exp\left(-\frac{\rho_1}{2t_k}\|u\|_2^2 + \frac{\sigma_n}{t_k} \mathsf{R}_n((u,0)_\delta;\mathbf{Z}) \right).$$

It remains to deal with the term $\int_{\mathbb{R}^{\|\delta\|_0}} M_{k,\delta}(\theta_\delta,\mathrm{d}v)\|v\|_2^2$. For clarity sake let's work is a slightly more general setting. Suppose that we have a density on $\mathbb{R}^d$, $d \geq 1$, that is proportional to $e^{-m(u)}$ for some function m of the form

$$m(u) \stackrel{\text{def}}{=} \frac{\rho}{2}\|u\|_2^2 + \ell(u),$$

for some bounded function ℓ. Let $q_\eta(u,\cdot)$ be the density of the proposal distribution $\mathbf{N}\left(\left(1 - \frac{\rho\eta^2}{2}\right)u, \eta^2 I_d\right)$, and define $\mathsf{R}(u) \stackrel{\text{def}}{=} \{v \in \mathbb{R}^d :\ \alpha(u,v) < 1\}$, where

$$\alpha(u,v) = \min\left(1, \frac{e^{-m(v)} q_\eta(v,u)}{e^{-m(u)} q_\eta(u,v)}\right).$$

Let L denote the resulting transition kernel on $\mathbb{R}^p$. We have

$$\begin{aligned}
\int_{\mathbb{R}^d} L(u,\mathrm{d}v)\|v\|_2^2 &= \|u\|_2^2 + \int \alpha(x,y)\left(\|v\|_2^2 - \|u\|_2^2\right) q_\eta(u,v)\mathrm{d}v \\
&= \|u\|_2^2 + \int_{\mathsf{R}(u)} \left[\frac{e^{-m(v)} q_\eta(v,u)}{e^{-m(u)} q_\eta(u,v)} - 1\right]\left(\|v\|_2^2 - \|u\|_2^2\right) q_\eta(u,v)\mathrm{d}v \\
&\quad + \int_{\mathbb{R}^d} \left(\|v\|_2^2 - \|u\|_2^2\right) q_\eta(u,v)\mathrm{d}v.
\end{aligned}$$

We can write

$$\|v\|_2^2 - \|u\|_2^2 = 2\langle u, v-u\rangle + \|v-u\|_2^2.$$

Integrating both sides, we get

$$\int_{\mathbb{R}^d} \left(\|v\|_2^2 - \|u\|_2^2\right) q_\eta(u,v)\mathrm{d}v = -\rho\eta^2\|u\|_2^2 + \frac{\rho^2\eta^4}{4}\|u\|_2^2 + \eta^2 d \leq -\frac{3}{4}\rho\eta^2\|u\|_2^2 + \eta^2 d, \quad \text{(S-25)}$$

by choosing η such that $\eta^2\rho \leq 1/2$. We also have

$$\frac{e^{-m(v)}q_\eta(v,u)}{e^{-m(u)}q_\eta(u,v)} = \exp\left(m(u) - m(v) - \frac{1}{2\eta^2}\left\|u - v + \frac{\rho\eta^2}{2}v\right\|^2 + \frac{1}{2\eta^2}\left\|v - u + \frac{\rho\eta^2}{2}u\right\|^2\right).$$

If $v \in \mathsf{R}(u)$, we necessarily have $\frac{e^{-m(v)}q_\eta(v,u)}{e^{-m(u)}q_\eta(u,v)} < 1$, which translates to:

$$m(v) - m(u) > -\frac{1}{2\eta^2}\left\|u - v + \frac{\rho\eta^2}{2}v\right\|^2 + \frac{1}{2\eta^2}\left\|v - u + \frac{\rho\eta^2}{2}u\right\|^2.$$

Noting that $m(u) = (\rho/2)\|u\|_2^2 + \ell(u)$, where ℓ is bounded by b_0, we infer that for $v \in \mathsf{R}(u)$,

$$\begin{aligned}\frac{\rho}{2}\left(\|u\|_2^2 - \|v\|_2^2\right) &\leq 2b_0 + \frac{1}{2\eta^2}\left\|u - v + \frac{\rho\eta^2}{2}v\right\|^2 - \frac{1}{2\eta^2}\left\|v - u + \frac{\rho\eta^2}{2}u\right\|^2 \\ &= 2b_0 + \frac{\rho^2\eta^2}{8}\left(\|v\|_2^2 - \|u\|_2^2\right) - \frac{\rho}{2}\langle v-u, v+u\rangle \\ &= 2b_0 + \frac{\rho^2\eta^2}{8}\left(\|v\|_2^2 - \|u\|_2^2\right) + \frac{\rho}{2}\left(\|u\|_2^2 - \|v\|_2^2\right).\end{aligned}$$

Hence, for $v \in \mathsf{R}(u)$,

$$\|u\|_2^2 - \|v\|_2^2 \leq \frac{16b_0}{\rho^2\eta^2},$$

which we use to write

$$\begin{aligned}&\int_{\mathsf{R}(u)} \left[\frac{e^{-m(v)}q_\eta(v,u)}{e^{-m(u)}q_\eta(u,v)} - 1\right]\left(\|v\|_2^2 - \|u\|_2^2\right) q_\eta(u,v)\mathrm{d}v \\ &\qquad = \int_{\mathsf{R}(u)} \left[1 - \frac{e^{-m(v)}q_\eta(v,u)}{e^{-m(u)}q_\eta(u,v)}\right]\left(\|u\|_2^2 - \|v\|_2^2\right) q_\eta(u,v)\mathrm{d}v \ \leq \frac{16b_0}{\rho^2\eta^2}. \quad \text{(S-26)}\end{aligned}$$

We combine (S-25)-(S-26) to conclude that

$$\int_{\mathbb{R}^d} L(u, \mathrm{d}v)\|v\|_2^2 \leq \|u\|_2^2 - \frac{3}{4}\rho\eta^2\|u\|_2^2 + \eta^2 d + \frac{16 b_0}{\rho^2\eta^2}.$$

Hence we can find b_1 (for instance $b_1 = 4\rho^{-1}(d + 16 b_0 \rho^{-2}\eta^{-4})$) such that for $\|u\|_2^2 > b_1$, it holds

$$\int_{\mathbb{R}^d} L(u, \mathrm{d}v)\|v\|_2^2 \leq \left(1 - \frac{\rho\eta^2}{2}\right)\|u\|_2^2.$$

This results applied to $M_{k,\delta}$ together with (S-24) implies that there exist $\lambda \in (0,1)$ (for instance $\lambda = \frac{\rho\eta^2}{2t_K}$), and $L < \infty$ such that

$$PV(\delta, \theta, k) \leq \lambda V(\delta, \theta, k), \quad \text{for all} (\delta, \theta, k) \notin \bar{\mathcal{C}}_L.$$

With similar (but simpler) calculations we check that

$$\sup_{(\delta,\theta,k)\in\bar{\mathcal{C}}_L} PV(\delta, \theta, k) \leq b,$$

for some finite constant b. This establishes the drift condition

$$PV(\delta, \theta, k) \leq \lambda V(\delta, \theta, k) + b\mathbf{1}_{\bar{\mathcal{C}}_L}(\delta, \theta, k), \quad \text{for all } (\delta, \theta, k) \in \mathcal{X}.$$

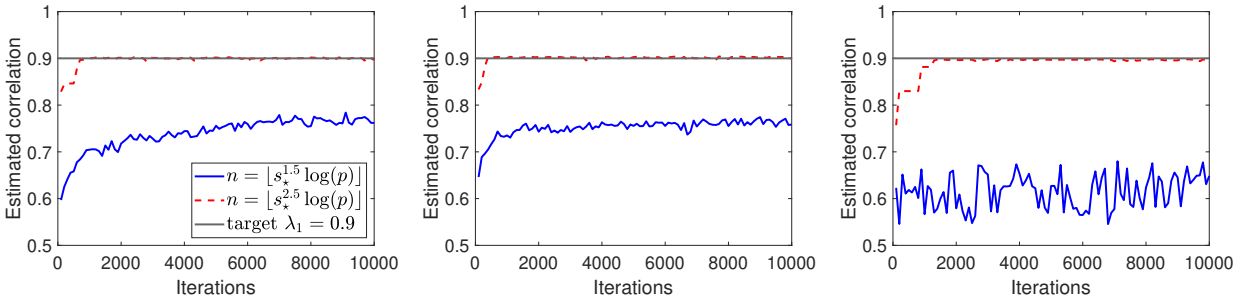
Hence the result.

S-4.3 Empirical studies of our algorithm

S-4.3.1 On sparse CCA computational barrier

It was conjectured by ([6]) that it is not possible to solve the sparse CCA problem in polynomial time at the statistical rate ϵ obtained in (13), in the data regime $n = o(s_\star^2 \log(p))$. The authors made a compelling argument for this conjecture by showing that any such estimator for the sparse CCA can be used to solve the planted clique problem in a regime where it is widely believed to be computationally intractable. Since our estimator achieves the rate ϵ under the weaker condition $n \geq C_0 s_\star \log(p)$, we have the opportunity to test empirically this conjecture.

In our simulation, we let $\Sigma_x \in \mathbb{R}^{(p/2)\times(p/2)}$ and $\Sigma_y \in \mathbb{R}^{(p/2)\times(p/2)}$ share the same structure, namely, a block diagonal matrix with five blocks, each of dimension $p/10 \times p/10$, where the (j, j')-th element of each block takes value $0.8^{|j-j'|}$. We let $\lambda_1 = 0.9, (v_{x\star})_j = (v_{y\star})_j = 1/\sqrt{3}$



(a) $p = 500$ (b) $p = 2000$ (c) $p = 5000$

Figure 1: Estimated canonical correlation along the MCMC iterations, averaged over 30 data repetitions for different values of dimension p and sample size n.

for $j \in \{1, 6, 11\}$, and $(v_{x\star})_j = (v_{y\star})_j = 0$ otherwise. Therefore, the true density level is $s_\star = 6$. For each $p \in \{500, 2000, 5000\}$, we generate data from the model described in Section 4.1 with two values of the sample size n, namely $\lceil s_\star^{1.5} \log(p) \rceil$ and $\lceil s_\star^{2.5} \log(p) \rceil$. We use the sample covariance matrices as estimators of Σ_x, Σ_y and Σ_{xy}, and set the scaling parameter $\sigma_n = 2n$ to construct the extended posterior distribution $\bar{\Pi}$ in (15). We sample from $\bar{\Pi}$ using Algorithm 2, with the set of temperatures $\{1, 1/0.9, 1/0.8, 1/0.7\}$. Since in this particular data model, the largest value of the (population) Rayleigh quotient is $\lambda_1 = 0.9$, proximity of the sample Rayleigh quotient $\mathsf{R}_n(\cdot; \mathbf{Z})$ to λ_1 along the MCMC iterations is a good empirical measure of mixing.

We run each MCMC sampler for $N = 10,000$ iterations, repeated 30 times (each time with a newly generated dataset). At each iteration time, we average the values of $\mathsf{R}_n(\cdot; \mathbf{Z})$ across the 30 repetitions. Fig. 1 shows the plot of the averaged sample Rayleigh quotient along iterations. The difference in behavior is striking. We clearly see that for all values of p, the sample Rayleigh quotient $\mathsf{R}_n(\cdot; \mathbf{Z})$ corresponding to $n = \lceil s_\star^{2.5} \log(p) \rceil$ quickly converges to the population Rayleigh quotient $\lambda_1 = 0.9$, whereas the one corresponding to $n = \lceil s_\star^{1.5} \log(p) \rceil$ fails to converge even after 10,000 iterations. This suggests that the condition $n \geq C_0 s_\star^2 \log(p)$ is indeed needed for the simulated tempering sampler to mix well, which appears to confirm the conjecture by ([6]).

S-4.3.2 Empirical mixing time of Algorithm 1

We investigate more carefully the mixing time of Algorithm 1 as a function of the dimension p, using the coupled chain approach of ([5, 9]) as described in Section S-4. We focus on a

data-rich setting where the sample size $n = p/2$. Now, let us describe the implementation details. We let Σ_x, Σ_y, $v_{x\star}$ and $v_{y\star}$ all have the same structures as in Section S-4.3.1 and set $\lambda_1 = 0.9$. We generate datasets from the model in Section 4.1 for each $p \in \{100, 200, \ldots, 5000\}$, with sample size $n = p/2$. The extended posterior distribution $\bar{\Pi}$ in (15) is constructed in the same way as in Section S-4.3.1, except with the set of temperatures $\{1, 1/0.9, 1/0.8, 1/0.7, 1/0.6\}$. We set the lag $L = p$ and the maximum iterations $N = 10p + 1000$. For each value of p, we repeat the simulation 50 times to estimate the distribution of the meeting time $\tau^{(L)}$ of the chain. More precisely, using $\varepsilon = 0.1$, we estimate the mixing time of the chain as the first iteration t for which the Monte Carlo estimate of the right hand side of (S-20) is less than ε. Fig. 2 below shows the plot of the mean of meeting times and the estimated mixing times as functions of p. The results suggest that Algorithm 2 has a mixing time that scales roughly linearly in the dimension p.

Remark S-4.1. *As far as we know, the existing literature on simulated tempering gives only general guidelines on choosing the temperatures ([8, 2]). The implementation of these guidelines remains challenging, and typically requires further adaptive MCMC methods ([12]). In our case, the Rayleigh quotient responds very well to temperature tuning, and in particular does not require high temperatures to mix well. As a result, we have chosen to maintain some very simple temperature scaling, and these work very well in the our experiments.*

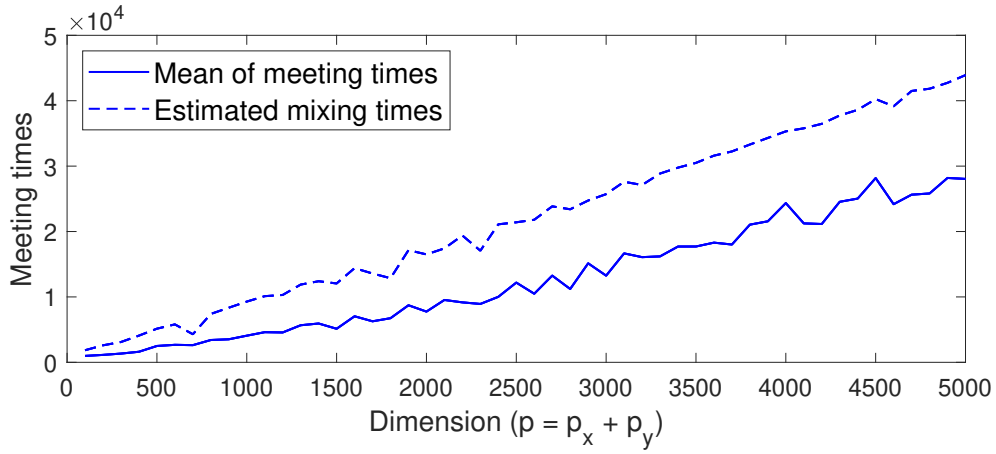


Figure 2: The mean of meeting times versus the estimated mixing times. The estimated mixing times are with respect to the total variation distance 0.1.

References

[1] Christophe Andrieu and Johannes Thoms. A tutorial on adaptive mcmc. *Statistics and Computing*, 18(4):343–373, 2008.

[2] Yves Atchadé, Gareth Roberts, and Jeffrey Rosenthal. Towards optimal scaling of metropolis-coupled markov chain monte carlo. *Statistics and Computing*, 21:555–568, 10 2011.

[3] Yves Atchadé and Jeffrey Rosenthal. On adaptive markov chain monte carlo algorithms. *Bernoulli*, 11, 10 2005.

[4] Yves F. Atchadé and Jun S. Liu. The wang-landau algorithm in general state spaces: Applications and convergence analysis. *Statistica Sinica*, 20(1):209–233, 2010.

[5] Niloy Biswas, Pierre E. Jacob, and Paul Vanetti. Estimating convergence of markov chains with l-lag couplings, 2019.

[6] Chao Gao, Zongming Ma, and Harrison H. Zhou. Sparse cca: Adaptive estimation and computational barriers. *Ann. Statist.*, 45(5):2074–2101, 10 2017.

[7] Rong Ge, Holden Lee, and Andrej Risteski. Simulated Tempering Langevin Monte Carlo II: An Improved Proof using Soft Markov Chain Decomposition. *arXiv e-prints*, page arXiv:1812.00793, November 2018.

[8] Charles J. Geyer and Elizabeth A. Thompson. Annealing markov chain monte carlo with applications to ancestral inference. *J. Amer. Stat. Assoc.*, 90(431):909–920, 1995.

[9] Pierre Jacob, John O'Leary, and Yves Atchadé. Unbiased markov chain monte carlo with couplings. *J. R. Stat. Soc. Ser. B*, 08 2017.

[10] Shengqiao Li. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics and Statistics*, 4(1):66–70, 2011.

[11] Jun S Liu. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2008.

[12] Btażej Miasojedow, Eric Moulines, and Matti Vihola. An adaptive parallel tempering algorithm. *Journal of Computational and Graphical Statistics*, 22(3):649–664, 2013.

[13] Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

[14] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[15] Vincent Q. Vu and Jing Lei. Minimax sparse principal subspace estimation in high dimensions. *Ann. Statist.*, 41(6):2905–2947, 12 2013.

[16] Dawn B. Woodard, Scott C. Schmidler, and Mark Huber. Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. *Ann. Appl. Probab.*, 19(2):617–640, 04 2009.