Supplementary to Nonparametric Tree Graphical Models via Kernel Embeddings

Le Song, Arthur Gretton, Carlos Guestrin

March 22, 2010

The supplementary material contains proofs of the main theorems (Section 1), and two additional experiments (Section 2): a reconstruction of camera orientation from images; and an additional set of document retrieval experiments, using a language graph constructed via the Chow-Liu algorithm.

1 Proofs

1.1 Preliminary results

Given any operator $A : \mathcal{G} \to \mathcal{F}$, the operator norm of A is written $\|A\|_2$, and its Hilbert-Schmidt norm (where defined) is

$$\|A\|_{HS}^2 := \sum_{i,j=1}^{\infty} \langle \varphi_j, A\phi_i \rangle_{\mathcal{F}}^2 ,$$

where the φ_i form a complete orthonormal system (CONS) for $\mathcal{F}$, and the ϕ_j form a CONS for $\mathcal{G}$. The set of Hilbert-Schmidt operators has the inner product

$$\langle A, B \rangle_{HS} = \sum_{i,j \geq 1} \langle A\phi_i, \varphi_j \rangle_{\mathcal{F}} \langle B\phi_i, \varphi_j \rangle_{\mathcal{F}}$$

We have defined the rank one operator $f \otimes g : \mathcal{G} \to \mathcal{F}$ such that $f \otimes g(h) = \langle g, h \rangle_{\mathcal{G}} f$.

It follows that

$$\langle f \otimes g, A \rangle_{HS} = \langle Ag, f \rangle_{\mathcal{F}} ,$$

and in particular,

$$\langle a \otimes b, u \otimes v \rangle_{HS} = \langle a, u \rangle_{\mathcal{F}} \langle b, v \rangle_{\mathcal{G}} .$$

We can extend this notation to higher order: for instance, given the product space $\mathcal{F}^n$ and functions $a_i \in \mathcal{F}$ and $b_i \in \mathcal{F}$ for $i \in \{1, \ldots, n\}$,

$$\left\langle \bigotimes_{i=1}^{n} a_i, \bigotimes_{i=1}^{n} b_i \right\rangle_{\mathcal{F}^n} = \prod_{i=1}^{n} \langle a_i, b_i \rangle_{\mathcal{F}} . \tag{1}$$

We use the result

$$A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}. \tag{2}$$

Further, following [3], we may define the empirical regularized correlation operator $\hat{\mathcal{V}}_{XY}$ such that

$$\hat{\mathcal{C}}_{XY} := \left(\hat{\mathcal{C}}_{XX} + \lambda_m I\right)^{1/2} \hat{\mathcal{V}}_{XY} \left(\hat{\mathcal{C}}_{YY} + \lambda_m I\right)^{1/2}. \tag{3}$$

where we have $\left\|\hat{\mathcal{V}}_{XY}\right\| \leq 1$.

1.2 Proof of Theorem 1

We now prove the result

$$\left\|\hat{\mathcal{U}}_{Y|X} - \mathcal{U}_{Y|X}\right\|_{HS} = O_p(\lambda_m^{\frac{1}{2}} + \lambda_m^{-\frac{3}{2}} m^{-\frac{1}{2}}). \tag{4}$$

We define a regularized population operator

$$\tilde{\mathcal{U}}_{Y|X} := \mathcal{C}_{YX} \left(\mathcal{C}_{XX} + \lambda_m I\right)^{-1}$$

and decompose (4) as

$$\left\|\hat{\mathcal{U}}_{Y|X} - \mathcal{U}_{Y|X}\right\|_{HS} \leq \left\|\hat{\mathcal{U}}_{Y|X} - \tilde{\mathcal{U}}_{Y|X}\right\|_{HS} + \left\|\mathcal{U}_{Y|X} - \tilde{\mathcal{U}}_{Y|X}\right\|_{HS}.$$

There are two parts to the proof. In the first part, we show convergence in probability of the first term in the above sum. In the second part, we demonstrate that as long as $\mathcal{C}_{YX}\mathcal{C}_{XX}^{-\frac{3}{2}}$ is Hilbert-Schmidt, the second term in the sum converges to zero as λ_m drops.

Part 1: We make the decomposition

$$\begin{aligned} & \left\|\mathcal{C}_{YX} \left(\mathcal{C}_{XX} + \lambda_m I\right)^{-1} - \hat{\mathcal{C}}_{YX} \left(\hat{\mathcal{C}}_{XX} + \lambda_m I\right)^{-1}\right\|_{HS} \\ \leq & \left\|\left(\mathcal{C}_{YX} - \hat{\mathcal{C}}_{YX}\right) \left(\mathcal{C}_{XX} + \lambda_m I\right)^{-1}\right\|_{HS} + \left\|\hat{\mathcal{C}}_{YX} \left[\left(\hat{\mathcal{C}}_{XX} + \lambda_m I\right)^{-1} - \left(\mathcal{C}_{XX} + \lambda_m I\right)^{-1}\right]\right\|_{HS}. \end{aligned}$$

The first term is bounded according to

$$\left\|\left(\mathcal{C}_{YX} - \hat{\mathcal{C}}_{YX}\right) \left(\mathcal{C}_{XX} + \lambda_m I\right)^{-1}\right\|_{HS} \leq \frac{1}{\lambda_m} \left\|\mathcal{C}_{YX} - \hat{\mathcal{C}}_{YX}\right\|_{HS},$$

and we know from [2, Lemma 5] that $\left\|\mathcal{C}_{YX} - \hat{\mathcal{C}}_{YX}\right\|_{HS} = O_p(1/\sqrt{m})$. For the

second term, we first substitute (2) and then (3) to obtain

$$\begin{aligned}
& \left\| \hat{\mathcal{C}}_{YX} \left[\left(\hat{\mathcal{C}}_{XX} + \lambda_m I \right)^{-1} - \left(\mathcal{C}_{XX} + \lambda_m I \right)^{-1} \right] \right\|_{HS} \\
= & \left\| \hat{\mathcal{C}}_{YX} \left(\hat{\mathcal{C}}_{XX} + \lambda_m I \right)^{-1} \left[\mathcal{C}_{XX} - \hat{\mathcal{C}}_{XX} \right] \left(\mathcal{C}_{XX} + \lambda_m I \right)^{-1} \right\|_{HS} \\
= & \left\| \left(\hat{\mathcal{C}}_{YY} + \lambda_m I \right)^{1/2} \hat{\mathcal{V}}_{XY} \left(\mathcal{C}_{XX} + \lambda_m I \right)^{-1/2} \left[\mathcal{C}_{XX} - \hat{\mathcal{C}}_{XX} \right] \left(\mathcal{C}_{XX} + \lambda_m I \right)^{-1} \right\|_{HS} \\
\leq & \frac{\left\| \left(\hat{\mathcal{C}}_{YY} + \lambda_m I \right)^{1/2} \right\|}{\lambda_m^{3/2}} \left\| \mathcal{C}_{XX} - \hat{\mathcal{C}}_{XX} \right\|_{HS} = O_p(\lambda_m^{-\frac{3}{2}} m^{-\frac{1}{2}}).
\end{aligned}$$

Part 2: $\left\| \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1} - \mathcal{C}_{YX} \left(\mathcal{C}_{XX} + \lambda_m I \right)^{-1} \right\|_{HS} = O(\lambda_m^{\frac{1}{2}})$.

Proof: We first expand the covariance operator $\mathcal{C}_{XX}$ in terms of the complete orthonormal system (CONS)

$$\mathcal{C}_{XX} = \sum_{i=1}^{\infty} \nu_i \varphi_i \otimes \varphi_i. \tag{5}$$

Then

$$\begin{aligned}
& \left\| \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1} - \mathcal{C}_{YX} \left(\mathcal{C}_{XX} + \lambda_m I \right)^{-1} \right\|_{HS}^2 \qquad (6) \\
= & \sum_{i,j=1}^{\infty} \left\langle \phi_j, \left(\mathcal{C}_{YX}\mathcal{C}_{XX}^{-1} - \mathcal{C}_{YX} \left(\mathcal{C}_{XX} + \lambda_m I \right)^{-1} \right) \varphi_i \right\rangle^2 \\
= & \sum_{i,j=1}^{\infty} \left\langle \phi_j, \mathcal{C}_{YX}\nu_i^{-1}\varphi_i - \mathcal{C}_{YX}(\lambda_m + \nu_i)^{-1}\varphi_i \right\rangle^2 \\
= & \sum_{i,j=1}^{\infty} \left(\frac{\lambda_m}{\nu_i + \lambda_m} \right)^2 \left\langle \phi_j, \mathcal{C}_{YX}\nu_i^{-1}\varphi_i \right\rangle^2 \\
= & \sum_{i,j=1}^{\infty} \left(\frac{\lambda_m}{\nu_i + \lambda_m} \right)^2 \left\langle \phi_j, \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}\varphi_i \right\rangle^2
\end{aligned}$$

Next, define

$$s_{ji} := \langle \phi_j, \mathcal{C}_{YX}\varphi_i \rangle$$

Assuming $\mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}$ is Hilbert-Schmidt, we have that

$$\sum_{i,j=1}^{\infty} \left\langle \phi_j, \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}\varphi_i \right\rangle^2 = \sum_{i,j=1}^{\infty} \frac{s_{ji}^2}{\nu_i^2} \quad \text{is finite.}$$

Furthermore,

$$\left(\frac{\lambda_m}{\nu_i + \lambda_m} \right)^2 = \left(\frac{1}{\frac{1}{\frac{\lambda_m}{\nu_i}} + \frac{1}{1}} \right)^2 \leq \left(\frac{1}{2} \sqrt{\frac{\lambda_m}{\nu_i}} \right)^2 = \frac{1}{4} \frac{\lambda_m}{\nu_i}$$

where we have used the arithmetic-geometric-harmonic means inequality. Therefore we need

$$\sum_{i,j=1}^{\infty} \frac{1}{4} \frac{\lambda_m}{\nu_i} \frac{s_{ji}^2}{\nu_i^2} \quad \text{to be finite.}$$

If we assume that

$$c := \sum_{i,j=1}^{\infty} \frac{1}{4} \frac{s_{ji}^2}{\nu_i^3} \quad \text{is finite,}$$

which corresponds to $\mathcal{C}_{YX}\mathcal{C}_{XX}^{-\frac{3}{2}}$ being Hilbert-Schmidt, then the squared norm difference in (6) will approach zero with rate $\lambda_m c$.

1.3 Proof of Theorem 2

We make a similar decomposition to the proof of Theorem 1, yielding

$$\begin{aligned}
& \left\| \left(\mathcal{C}_{YY} + \lambda_m I\right)^{-1} \mathcal{C}_{YX} \left(\mathcal{C}_{XX} + \lambda_m I\right)^{-1} - \left(\hat{\mathcal{C}}_{YY} + \lambda_m I\right)^{-1} \hat{\mathcal{C}}_{YX} \left(\hat{\mathcal{C}}_{XX} + \lambda_m I\right)^{-1} \right\|_{HS} \\
\leq & \left\| \left[\left(\mathcal{C}_{YY} + \lambda_m I\right)^{-1} - \left(\hat{\mathcal{C}}_{YY} + \lambda_m I\right)^{-1} \right] \mathcal{C}_{YX} \left(\mathcal{C}_{XX} + \lambda_m I\right)^{-1} \right\|_{HS} \\
& + \left\| \left(\hat{\mathcal{C}}_{YY} + \lambda_m I\right)^{-1} \left(\mathcal{C}_{YX} - \hat{\mathcal{C}}_{YX}\right) \left(\mathcal{C}_{XX} + \lambda_m I\right)^{-1} \right\|_{HS} \\
& + \left\| \left(\hat{\mathcal{C}}_{YY} + \lambda_m I\right)^{-1} \hat{\mathcal{C}}_{YX} \left[\left(\mathcal{C}_{XX} + \lambda_m I\right)^{-1} - \left(\hat{\mathcal{C}}_{XX} + \lambda_m I\right)^{-1} \right] \right\|_{HS}.
\end{aligned}$$

The first term is bounded according to

$$\begin{aligned}
& \left\| \left[\left(\mathcal{C}_{YY} + \lambda_m I\right)^{-1} - \left(\hat{\mathcal{C}}_{YY} + \lambda_m I\right)^{-1} \right] \mathcal{C}_{YX} \left(\mathcal{C}_{XX} + \lambda_m I\right)^{-1} \right\|_{HS} \\
= & \left\| \left(\hat{\mathcal{C}}_{YY} + \lambda_m I\right)^{-1} \left[\mathcal{C}_{YY} - \hat{\mathcal{C}}_{YY}\right] \left(\mathcal{C}_{YY} + \lambda_m I\right)^{-1} \mathcal{C}_{YX} \left(\mathcal{C}_{XX} + \lambda_m I\right)^{-1} \right\|_{HS} \\
\leq & \left\| \left(\hat{\mathcal{C}}_{YY} + \lambda_m I\right)^{-1} \left[\mathcal{C}_{YY} - \hat{\mathcal{C}}_{YY}\right] \left(\mathcal{C}_{YY} + \lambda_m I\right)^{-1/2} V_{XY} \left(\mathcal{C}_{XX} + \lambda_m I\right)^{-1/2} \right\|_{HS} \\
\leq & \frac{\left\| \mathcal{C}_{YY} - \hat{\mathcal{C}}_{YY} \right\|_{HS}}{\lambda_m^2} = O_p(\lambda_m^{-2} m^{-\frac{1}{2}}).
\end{aligned}$$

The third term follows similar reasoning. The second term is bounded according to

$$\left\| (\hat{\mathcal{C}}_{YY} + \lambda_m I)^{-1} (\mathcal{C}_{YX} - \hat{\mathcal{C}}_{YX}) \left(\mathcal{C}_{XX} + \lambda_m I\right)^{-1} \right\|_{HS} \leq \frac{\left\| \mathcal{C}_{YX} - \hat{\mathcal{C}}_{YX} \right\|_{HS}}{\lambda_m^2} = O_p(\lambda_m^{-2} m^{-\frac{1}{2}}).$$

Convergence in probability of the three terms follows from the convergence of each of $\left\| \mathcal{C}_{YY} - \hat{\mathcal{C}}_{YY} \right\|_{HS}$, $\left\| \mathcal{C}_{YX} - \hat{\mathcal{C}}_{YX} \right\|_{HS}$, and $\left\| \mathcal{C}_{XX} - \hat{\mathcal{C}}_{XX} \right\|_{HS}$, as in the proof of Theorem 1.

We next address the convergence of

$$\left\|\left(\mathcal{C}_{YY}+\lambda_m I\right)^{-1}\mathcal{C}_{YX}\left(\mathcal{C}_{XX}+\lambda_m I\right)^{-1}-\mathcal{C}_{YY}^{-1}C_{XY}\mathcal{C}_{XX}^{-1}\right\|_{HS}.$$

for λ_m approaching zero. We use the earlier decomposition of $\mathcal{C}_{XX}$ in terms of its eigenfunctions φ_i from (5), and further require that ϕ_i be the eigenfunctions of $\mathcal{C}_{YY}$,

$$\mathcal{C}_{YY}:=\sum_{i=1}^{\infty}\gamma_i\phi_i\otimes\phi_i.$$

Thus

$$\begin{aligned}
&\left\|\mathcal{C}_{YY}^{-1}C_{XY}\mathcal{C}_{XX}^{-1}-\left(\mathcal{C}_{YY}+\lambda_m I\right)^{-1}\mathcal{C}_{YX}\left(\mathcal{C}_{XX}+\lambda_m I\right)^{-1}\right\|_{HS}^2\\
=&\sum_{i,j=1}^{\infty}\left\langle\phi_j,\left(\mathcal{C}_{YY}^{-1}C_{XY}\mathcal{C}_{XX}^{-1}-\left(\mathcal{C}_{YY}+\lambda_m I\right)^{-1}\mathcal{C}_{YX}\left(\mathcal{C}_{XX}+\lambda_m I\right)^{-1}\right)\varphi_i\right\rangle^2\\
=&\sum_{i,j=1}^{\infty}\left\langle\phi_j,\mathcal{C}_{YY}^{-1}C_{XY}\nu_i^{-1}\varphi_i-\left(\mathcal{C}_{YY}+\lambda_m I\right)^{-1}\mathcal{C}_{YX}(\nu_i+\lambda_m)^{-1}\varphi_i\right\rangle^2\\
=&\sum_{i,j=1}^{\infty}\left\langle\phi_j,C_{XY}\left(\gamma_j\nu_i\right)^{-1}\varphi_i-\mathcal{C}_{YX}(\nu_i+\lambda_m)^{-1}(\gamma_j+\lambda_m)^{-1}\varphi_i\right\rangle^2\\
=&\sum_{i,j=1}^{\infty}\left(\frac{\lambda_m^2+\gamma_j\lambda_m+\nu_i\lambda_m}{\left(\nu_i+\lambda_m\right)\left(\gamma_j+\lambda_m\right)}\right)^2\left\langle\phi_j,\mathcal{C}_{YY}^{-1}C_{XY}\mathcal{C}_{XX}^{-1}\varphi_i\right\rangle^2.
\end{aligned}$$

Furthermore, we have

$$\left(\frac{\lambda_m^2+\gamma_j\lambda_m+\nu_i\lambda_m}{\nu_i\gamma_j+\lambda_m^2+\gamma_j\lambda_m+\nu_i\lambda_m}\right)^2\leq\frac{1}{4}\left(\frac{\lambda_m^2+\gamma_j\lambda_m+\nu_i\lambda_m}{\nu_i\gamma_j}\right),$$

where we again use the arithmetic-geometric-harmonic mean inequality. Assuming $\lambda_m\ll\gamma_1$ and $\lambda_m\ll\nu_1$, it follows that

$$\lambda_m^2<\gamma_1\lambda_m+\nu_1\lambda_m,$$

and thus

$$\frac{1}{4}\left(\frac{\lambda_m^2+\gamma_j\lambda_m+\nu_i\lambda_m}{\nu_i\gamma_j}\right)<\frac{1}{2}\left(\frac{\gamma_1\lambda_m+\nu_1\lambda_m}{\nu_i\gamma_j}\right).$$

We therefore require the finiteness of

$$\sum_{i,j=1}^{\infty}\frac{\lambda_m}{2}\left(\frac{\nu_1+\gamma_1}{\nu_i\gamma_j}\right)\frac{s_{ij}^2}{\nu_i^2\gamma_j^2}<\frac{\lambda_m\left(\nu_1+\gamma_1\right)}{2}\sum_{i,j=1}^{\infty}\frac{s_{ij}^2}{\nu_i^3\gamma_j^3}.$$

This is equivalent to requiring that $\mathcal{C}_{YY}^{-\frac{3}{2}}\mathcal{C}_{YX}\mathcal{C}_{XX}^{-\frac{3}{2}}$ be Hilbert-Schmidt as a condition of convergence.

1.4 Proof of Theorem 3

Our bound is in terms of the following constants:

$$R_m = \max_{(s,t)\in\mathcal{E}} \frac{\|M_{ts}\|_{\mathcal{H}_t}}{\|m_{ts}\|_{\mathcal{F}}} \tag{7}$$

$$R_B = \max_{t\in\mathcal{V}} \frac{\|M_{ts}\otimes m_{st}\|_{\mathcal{H}_t\otimes\mathcal{F}}}{\|B_t\|_{\mathcal{F}}} \tag{8}$$

$$R_{\mathcal{L}} = \max_{(s,t)\in\mathcal{E}} \sup_{x_t\in\mathcal{X}} \|f_{x_t}\|_{\mathcal{F}}^{-1} \tag{9}$$

$$R = \max\{R_m, R_B, R_{\mathcal{L}}\} \tag{10}$$

where R_m is the maximal ratio of the RKHS norm of the pre-message to that of the message, R_B is maximal ratio of the RKHS norm of the pre-belief to that of the belief, and $R_{\mathcal{L}}$ is the maximal inverse of the RKHS norm of f_{x_t}. R is the largest of these three quantities. R_m and R_B quantify the degree of smoothing of the RKHS function after message propagation, while $R_{\mathcal{L}}$ quantifies the smoothness of the RKHS function f_{x_t} itself. Under our assumption that $0 \leq k(x,x') = \langle\varphi(x),\varphi(x')\rangle_{\mathcal{F}} \leq 1$, we have $\|\hat{\mathcal{U}}_{X_t^{d_t-1}|X_s}\|_2 \leq 1$ and $\left\|\mathcal{C}_{X_s^{d_s}X_s}\right\|_2 \leq 1$.

Proof We first bound the difference between the true message $m_{ts} = M_{ts}^\top \mathcal{U}_{X_t|X_s}$ and the message produced by propagating the true "pre-message" through the estimated embedding operator $\tilde{m}_{ts} := M_{ts}^\top \hat{\mathcal{U}}_{X_t^{d_t-1}|X_s}$:

$$\begin{aligned}
\frac{\|\tilde{m}_{ts} - m_{ts}\|_{\mathcal{F}}}{\|m_{ts}\|_{\mathcal{F}}} =& \frac{\left\|M_{ts}^\top \mathcal{U}_{X_t^{d_t-1}|X_s} - M_{ts}^\top \hat{\mathcal{U}}_{X_t^{d_t-1}|X_s}\right\|_{\mathcal{F}}}{\|m_{ts}\|_{\mathcal{F}}} \\
&\leq \frac{\|M_{ts}\|_{\mathcal{H}_t}}{\|m_{ts}\|} \left\|\mathcal{U}_{X_t^{d_t-1}|X_s} - \hat{\mathcal{U}}_{X_t^{d_t-1}X_s}\right\|_{HS} \\
&\leq RC\left(\tfrac{\delta}{2(n-1)}\right)\lambda^{-2}m^{-\frac{1}{2}} =: \epsilon
\end{aligned} \tag{11}$$

with probability at least $1-\delta$ simultaneously for all $2(n-1)$ messages, using the union bound. The first inequality follows from $\|\mathcal{T}a\|_{\mathcal{F}} \leq \|\mathcal{T}\|_2 \|a\|_{\mathcal{F}}$, and the relation between the spectral norm and Hilbert-Schmidt norm of operators, *i.e.* $\|\mathcal{T}\|_2 \leq \|\mathcal{T}\|_{HS}$. We then have

$$\tilde{m}_{ts} \in m_{ts} + v\cdot\epsilon\|m_{ts}\|_{\mathcal{F}}, \quad \|v\|_{\mathcal{F}} \leq 1 \tag{12}$$

Note that $\tilde{m}_{ts}$ is different from the estimated message $\hat{m}_{ts}(x_s) := \hat{M}_{ts}^\top \hat{\mathcal{U}}_{X_t^{d_t-1}|X_s}\varphi(x_s)$, where both the pre-message and the conditional embedding operator are estimated. Next, we bound

$$\begin{aligned}
\frac{\|\hat{m}_{ts} - m_{ts}\|_{\mathcal{F}}}{\|m_{ts}\|_{\mathcal{F}}} &\leq \frac{\|\hat{m}_{ts} - \tilde{m}_{ts}\|_{\mathcal{F}}}{\|m_{ts}\|_{\mathcal{F}}} + \frac{\|\tilde{m}_{ts} - m_{ts}\|_{\mathcal{F}}}{\|m_{ts}\|_{\mathcal{F}}} \\
&\leq \frac{\left\|\hat{M}_{ts}^\top \hat{\mathcal{U}}_{X_t^{d_t-1}|X_s} - M_{ts}^\top \hat{\mathcal{U}}_{X_t^{d_t-1}|X_s}\right\|_{\mathcal{F}}}{\|m_{ts}\|_{\mathcal{F}}} + \epsilon \\
&\leq \frac{\left\|\hat{M}_{ts} - M_{ts}\right\|_{\mathcal{H}_t}}{\|m_{ts}\|_{\mathcal{F}}} + \epsilon
\end{aligned} \tag{13}$$

where we use $\|\hat{\mathcal{U}}_{X_t^{d_t-1}|X_s}\|_2 \leq 1$. Furthermore, we have:

$$\begin{aligned}
&\frac{\left\|\hat{M}_{ts} - M_{ts}\right\|_{\mathcal{H}_t}}{\|m_{ts}\|_{\mathcal{F}}} \\
=&\frac{\|\bigotimes_u (m_{ut} + v_u \cdot \epsilon_u \|m_{ut}\|) - \bigotimes_u m_{ut}\|_{\mathcal{H}_t}}{\|m_{ts}\|_{\mathcal{F}}} \\
=&\frac{\|M_{ts}\|_{\mathcal{H}_t}}{\|m_{ts}\|_{\mathcal{F}}} \left\|\bigotimes_u (w_u + v_u \cdot \epsilon_u) - \bigotimes_u w_u\right\|_{\mathcal{H}_t} \\
&\left(\|M_{ts}\|_{\mathcal{H}_t} = \prod_u \|m_{ut}\|_{\mathcal{F}} \text{ and } \|w_u\|_{\mathcal{F}} = 1\right) \\
\leq& R \left\|\bigotimes_u w_u(1+\epsilon_u) - \bigotimes_u w_u\right\|_{\mathcal{H}_t} \\
\leq& R\left(\prod_u (1+\epsilon_u) - 1\right) \\
=& R\left(\sum_u \epsilon_u + \sum_{u,u'} O(\epsilon_u \epsilon_{u'})\right)
\end{aligned} \tag{14}$$

We can then prove by induction that

$$\frac{\|\hat{m}_{ts} - m_{ts}\|_{\mathcal{F}}}{\|m_{ts}\|_{\mathcal{F}}} \leq \epsilon \sum_{i \in \mathcal{T}_t} R^{h_i} + O(\epsilon^2) =: \epsilon_t \tag{15}$$

where $\mathcal{T}_t$ is the subtree induced by node t when it sends a message to s. For a node i in the subtree $\mathcal{T}_t$, h_i denotes the depth of this node. The root node of the subtree $\mathcal{T}_t$, *i.e.* node t, starts with depth 0, *i.e.* $h_t = 0$.

For a leaf node, the subtree $\mathcal{T}_t$ contains a single node, and $m_{ts} = f_{x_t}$. We have

$$\frac{\left\|\hat{f}_{x_t} - f_{x_t}\right\|_{\mathcal{F}}}{\|f_{x_t}\|_{\mathcal{F}}} \leq \frac{\left\|\hat{\mathcal{A}}_{ts} - \mathcal{A}_{ts}\right\|_{HS}}{\|f_{x_t}\|_{\mathcal{F}}} \leq \epsilon. \tag{16}$$

Assume that (15) holds for all messages coming into node t. Combining (13) and (14),

$$\begin{aligned}
\frac{\|\hat{m}_{ts} - m_{ts}\|_{\mathcal{F}}}{\|m_{ts}\|_{\mathcal{F}}} &\leq \epsilon \sum_u \sum_{i \in \mathcal{T}_u} R^{h_i+1} + O(\epsilon^2) \\
&= \epsilon \sum_{j \in \mathcal{T}_t} R^{h_j} + O(\epsilon^2)
\end{aligned} \tag{17}$$

where in the last equality we have grown the tree by one level. Applying a similar argument to the final belief B_s and using $\left\|\mathcal{C}_{X_s^{d_s} X_s}\right\|_2 \leq 1$, we complete the proof. ■

2 Additional experiments

Finding camera rotations: We apply NTGM to to a computer vision problem as in [5]. We try to determine the camera orientation based on the images it observes. In this setting, the camera focal point is fixed at a position and traces out a smooth path of rotations while making observations. The dataset is generated by POVRAY[1] which renders images observed by the camera. The virtual scene is a rectangular-shaped room with a ceiling light and two pieces of furniture. The images exhibit complex lighting effects such as shadows, inter-reflections, and global illumination, all of which make determining the camera rotation difficult especially for noisy cases.

The sequence of image observations contains 3600 frames, and we use the first 1800 frames for training and the remaining 1800 frames for testing. The dynamics governing the camera rotation is a piece-wise smooth random walk. This is an unconventional graphical model in that the camera state is a rotation matrix R from $SO(3)$; and the observations are images which are high dimensional spaces with correlation between pixel values. The graph structure for this problem is a caterpillar tree in Figure 1(b), and one performs online inference.

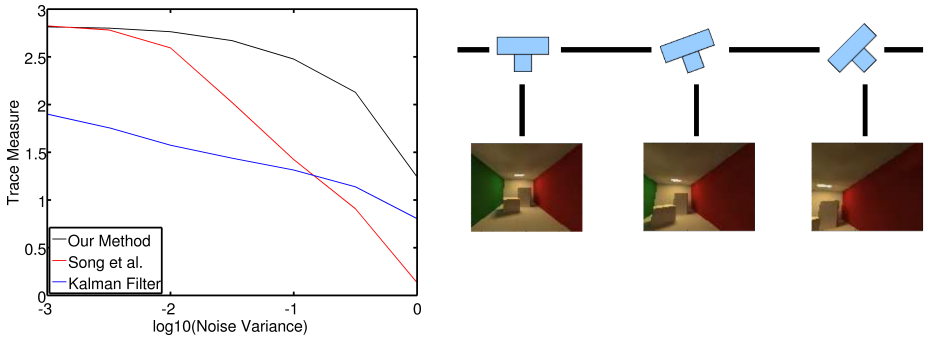
We flatten each image to a vector, and apply a Gaussian RBF kernel. The bandwidth parameter of the kernel is fixed using the median distance between image vectors. We use a Gaussian RBF kernel between two rotations R and $\tilde{R}$, *i.e.*, $k(R, \tilde{R}) := \exp(-\sigma\|R - \tilde{R}\|^2)$. Using this kernel, we find the most probable camera rotation matrix by maximizing the belief $B(R)$ over the rotation group [1].

We compare our method to a Kalman filter and the method of [5]. For the Kalman filter, we used the quaternion corresponding to a rotation matrix R as the state and the image vectors as the observations. We learn the model parameters of the linear dynamical system using linear regression. In Song et al., an approximation algorithm is used for aggregating dynamical system history and the current image observation. We expect NTGM which incorporates both information in a principled way should outperform the method by [5]. We use $\mathrm{tr}(R^\top \hat{R})$ between the true rotation R and the estimated one $\hat{R}$ as performance measure (this measure ranges between $[-1, 3]$, and the larger the better performance).

We add zero mean Gaussian white noise to the images and study the performance scaling of the three methods as we increase the noise variance. We observe that the performance of NTGM degrades more gracefully than the other two methods (Figure 1(a)). For large noise, Kalman filter overtakes the method proposed by [5]. In this setting, the images are very noisy, and the dynamics become the key to determine the camera orientation. In this regime, NTGM significantly outperforms the other two methods, with 40% higher trace measure.

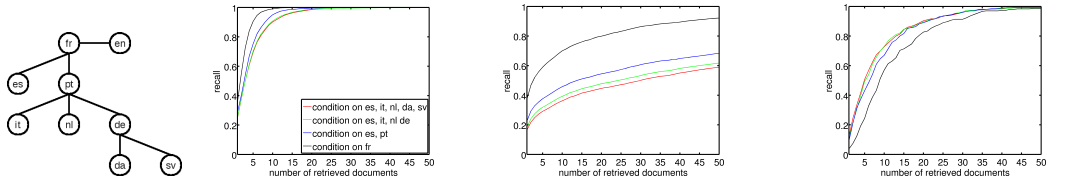
Additional cross-language document retrieval experiment: We obtained a graphical model on languages by applying the Chow-Liu algorithm,

[1] www.povray.org



(a) (b)

Figure 1: Performance of different methods vs observation noise, camera rotation problem.



(h) Tree 2 (i) NTGM (j) Bilingual Topic Model (k) Normalized File Size

Figure 2: (a) A graphical model for cross-language document retrieval, obtained via Chow-Liu with the HSIC dependence measure. The target document was in English. (f,g,h) The recall score for NTGM, bilingual topic model and normalized file size method for retrieval conditioned on document observations from other languages.

using the Hilbert-Schmidt Independence Criterion (HSIC) [4] for the required statistical dependence measure (applying the same kernels that were used in our inference algorithm). Our goal was to retrieve English documents conditioned on documents from other languages. Besides the different graph structure, all remaining experimental settings were identical to those of the linguistic similarity tree experiments (Figure 2(e) in the main document). Results are shown in Figure 2, and are qualitatively similar to the cross-language retrieval results using the linguistic similarity tree (Figures 2(f,g,h) in the main document).

References

[1] T. Abrudan, J. Eriksson, and V. Koivunen. Steepest descent algorithms for optimization under unitary matrix constraint. *IEEE SP*, 56(3), 2008.

[2] K. Fukumizu, F. Bach, and A. Gretton. Statistical consistency of kernel canonical correlation analysis. *JMLR*, 8:361–383, 2007.

[3] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems 20*, pages 489–496, Cambridge, MA, 2008. MIT Press.

[4] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *ALT 16*, pages 63–78, 2005.

[5] L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *ICML*, 2009.