
Supplement of "Detecting non-causal artifacts in multivariate linear regression models"

Dominik Janzing [1] **Bernhard Schölkopf** [2]

1. Proofs

1.1. Proof of Lemma 1

We first write Φ as $\Phi(v) = g(Av)Av$, with $g(w) := 1/\|w\|$. Let $t \mapsto s(t)$ be some curve on the unit sphere S^{d-1} and $\tilde{s}(t) := \Phi(s(t))$ its image and assume $v = s(0)$ and $\tilde{v} = \tilde{s}(0)$. Then we have

$$\begin{aligned}\frac{d}{dt}\Phi(s(t)) &= \langle \nabla g(As(t)), As'(t)\rangle As(t) \\ &\quad + g(As(t))As'(t),\end{aligned}$$

with $\nabla g(w) = -w/\|w\|^3$. Hence we obtain

$$\begin{aligned}&\frac{d}{dt}\Phi(s(t)) \qquad (1)\\ &= \frac{-1}{\|As(t)\|^3}\langle As(t), As'(t)\rangle As(t) + g(As(t))As'(t) \\ &= g(As(t))\left(As'(t) - \tilde{s}(t)\tilde{s}(t)^T As'(t)\right) \\ &= g(As(t))\left(\mathbf{1} - \tilde{s}(t)\tilde{s}(t)^T\right) As'(t), \qquad (2)\end{aligned}$$

where we have used $\tilde{s}(t) = As(t)/\|As(t)\|$. Note that the matrix $\mathbf{1} - \tilde{s}(t)\tilde{s}(t)^T$ projects $As'(t)$ onto the space orthogonal to $\tilde{s}(t)$, that is, the tangent space of the surface of the sphere at $\tilde{s}(t)$. Further, the matrix

$$g(As(t))\left(\mathbf{1} - \tilde{s}(t)\tilde{s}(t)^T\right) A$$

maps each tangent vector $s'(t)$ at $s(t)$ (for any curve s) to the corresponding tangent vector $\tilde{s}'(t)$ at $\tilde{s}(t)$. It thus describes the Jacobian $D\Phi$ mapping between tangent spaces $T_{s(t)}$ and $T_{\tilde{s}(t)}$ of the sphere at $s(t)$ and $\tilde{s}(t)$, respectively. Let $e_1, \dots, e_{d-1}$ and $\tilde{e}_1, \dots, \tilde{e}_{d-1}$ be orthonormal bases of T_v and $T_{\tilde{v}}$, respectively (that is, bases of $v^\perp$ and $\tilde{v}^\perp$, respectively). If we set $U_v := (e_1, \dots, e_{d-1})$ and $U_{\tilde{v}} := (\tilde{e}_1, \dots, \tilde{e}_{d-1})$, the matrix representation of the Jacobian $D\Phi$ with respect to these bases reads

$$\widehat{D\Phi}(v) := g(Av)U_{\tilde{v}}^T A U_v.$$

We then have

$$\det \widehat{D\Phi}(v) = g(Av)^{d-1}\det(U_{\tilde{v}}^T A U_v).$$

For later use, we also observe that multiplying the equation $\tilde{v} = Av/\|Av\|$ with A^{-1} and taking the norm on both sides yields

$$1/\|Av\| = \|A^{-1}\tilde{v}\|. \qquad (3)$$

For the probability density we thus obtain

$$\begin{aligned}p(\tilde{v}) &= |\det \widehat{D\Phi}(\Phi^{-1}(\tilde{v}))|^{-1} \\ &= \left(\|A^{-1}\tilde{v}\|^{d-1}|\det(U_{\tilde{v}}^T A U_v)|\right)^{-1} \\ &= \left(\|A^{-1}\tilde{v}\|^{d-1}|\det(\tilde{A})|\right)^{-1}, \qquad (4)\end{aligned}$$

with the abbreviation $\tilde{A} := U_{\tilde{v}}^T A U_v$. Let us now define the orthogonal $d \times d$ matrices

$$W_v := (U_v, v) \quad \text{and} \quad (U_{\tilde{v}}, \tilde{v}).$$

Then we define $A' := W_{\tilde{v}}^T A W_v$, which implies $|\det(A')| = |\det(A)|$. A' can be written as

$$A' = \begin{pmatrix} \tilde{A} & 0 \\ w & \|Av\| \end{pmatrix},$$

where w is some $1 \times (d-1)$-matrix. Hence we obtain

$$\det(A') = \det(\tilde{A})\|Av\| = \frac{\det(\tilde{A})}{\|A^{-1}\tilde{v}\|},$$

where we have used also (3). We can thus rewrite (4) as

$$p(\tilde{v}) = \frac{1}{|\det(A)|\|A^{-1}\tilde{v}\|^d}.$$

1.2. Proof of Theorem 3

By definition, $p_{\theta'}$ is obtained by applying the map $\sqrt{R'_\theta}$ to vectors drawn from a rotation invariant distribution with renormalizing it later. Without loss of generality, let all the matrices R_θ be diagonal with eigenvalues $f_j(\theta)$ (note that they commute). Let v be generated by drawing each entry v_j from $\mathcal{N}(0,1)$. We can then compute the entries of $\tilde{v}$ by

$$\tilde{v}_j := \frac{1}{\sum_{i=1}^d f_j(\theta')v_j^2}\sqrt{f_j(\theta')}v_j.$$

---

[1]Amazon Development Center, Tübingen, Germany [2]Max Planck Institute for Intelligent Systems, Tübingen, Germany. This work has been done at the MPI before DJ joined Amazon. Correspondence to: Dominik Janzing <janzind@amazon.com>.

Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

Rewriting (10) in terms of v_j instead of $\tilde{v}$ yields

$$\log p_\theta(\tilde{v}) = -\frac{1}{2}\left\{\log \frac{1}{d}\sum_{j=1}^{d} f_j(\theta')f_j(\theta)^{-1}v_j^2 - \log \frac{1}{d}\sum_{j=1}^{d} f_j(\theta')v_j^2\right\} + \frac{1}{2}\log\det R_\theta .$$

Since each v_j^2 is an independent squared standard Gaussian it has expectation 1 and variance 2. Therefore, the random variable

$$\frac{1}{d}\sum_{j=1}^{d} f_j(\theta')f_j(\theta)^{-1}v_j^2$$

has mean $\tau(R_\theta R_{\theta'}^{-1})$ and variance

$$\frac{2}{d^2}\sum_{j=1}^{d} f_j(\theta')^2 f_j(\theta)^{-2}.$$

Due to Chebychev's inequality we have

$$\left|\frac{1}{d}\sum_{j=1}^{d} f_j(\theta')f_j(\theta)^{-1}v_j^2 - \tau(R_{\theta'}R_\theta^{-1})\right| \leq \delta,$$

with probability $1 - \frac{2}{d^2}\sum_{j=1}^{d} f_j(\theta')^2 f_j(\theta)^{-2}/\delta^2 = 1 - \frac{2}{d}\tau(R_{\theta'}^2 R_\theta^{-2})/\delta^2$. Likewise,

$$\left|\frac{1}{d}\sum_{j=1}^{d} f_j(\theta')v_j^2 - \tau(R_{\theta'})\right| \leq \delta,$$

with probability $1 - \frac{2}{d}\tau(R_{\theta'}^2)/\delta^2$. Since $|\log(x+\rho) - \log x| \leq 2\rho/x$ for sufficiently small ρ, we can ensure that

$$\left|\log \frac{1}{d}\sum_{j=1}^{d} f_j(\theta')f_j(\theta)^{-1}v_j^2 - \log\tau(R_{\theta'}R_\theta^{-1})\right| \leq \epsilon, \quad (5)$$

by choosing $\delta \leq \epsilon/(2\tau(R_{\theta'}R_\theta^{-1}))$. Likewise, we can achieve that

$$\left|\log \frac{1}{d}\sum_{j=1}^{d} f_j(\theta')v_j^2 - \log\tau(R_{\theta'})\right| \leq \epsilon, \quad (6)$$

if $\delta \leq \epsilon/(2\tau(R_{\theta'}))$. Thus, both inequalities (5) and (6) together hold with probability at least

$$1 - \frac{8}{d\epsilon^2}\left[\tau(R_{\theta'}^2 R_\theta^{-2})\tau(R_{\theta'}R_\theta^{-1})^2 + \tau(R_{\theta'}^2)\tau(R_{\theta'})^2\right].$$