Supplementary Material: The Gaussian Process Autoregressive Regression Model (GPAR)

James Requeima[12] **Will Tebbutt**[12] **Wessel Bruinsma**[12] **Richard E. Turner**[1]
[1]University of Cambridge and [2]Invenia Labs, Cambridge, UK
{jrr41, wct23, wpb23, ret26}@cam.ac.uk

A Conditional Independence in Figure 2b

In this section, we prove the key conditional independence in Figure 2b that makes GPAR work:

Theorem 1. Let a set of observations $\mathcal{D}$ be closed downwards. Then $y_i \perp \mathcal{D}_{i+1:M} \,|\, \mathcal{D}_{1:i}$, where $\mathcal{D}_{1:i}$ are the observations for outputs $1, \ldots, i$ and $\mathcal{D}_{i+1:M}$ for $i+1, \ldots, M$.[1]

To begin with, we review some basic notions concerning graphical models. Let a path be a sequence of nodes $v_1, \ldots, v_n$ from some directed graph G where, for each v_i, G either contains an edge from v_i to v_{i+1}, or an edge from v_{i+1} to v_i. If G contains an edge from a to b, then write $a \to b$ to mean the two-node path in which node b follows node a. Similarly, if G contains an edge from b to a, then write $a \leftarrow b$ to mean the two-node path in which b follows a. Write $a \rightleftharpoons b$ to mean either $a \to b$ or $b \leftarrow a$.

Definition 1 (Active Path (Definition 3.6 from Koller and Friedman (2009)))**.** Let $\mathcal{P} = v_1 \rightleftharpoons \cdots \rightleftharpoons v_n$ be a path in a graphical model. Let Z be a subset of the variables from the graphical model. Then, call $\mathcal{P}$ active given Z if (1) for every v-structure $v_{i-1} \to v_i \leftarrow v_{i+1}$ in $\mathcal{P}$, v_i or a descendant of v_i is in Z; and (2) no other node in $\mathcal{P}$ is in Z.

Definition 2 (d-Separation (Definition 3.7 from Koller and Friedman (2009)))**.** Let X, Y, and Z be three sets of nodes from a graphical model. Then, call X and Y d-separated given Z if no path between any $x \in X$ and $y \in Y$ is active given Z.

Theorem 2 (d-Separation Implies Conditional Independence (Theorem 3.3 from Koller and Friedman (2009)))**.** Let X, Y, and Z be three sets of nodes from a graphical model. If X and Y are d-separated given Z, then $X \perp Y \,|\, Z$.

[1] $\mathcal{D}_{1:i} = \{y_j^{(n)} \in \mathcal{D} : j \leq i, n \leq N\}$ and $\mathcal{D}_{i+1:M} = \{y_j^{(n)} \in \mathcal{D} : j > i, n \leq N\}$.

Define the *layer* of a node in Figure 2b to be

$$\operatorname{layer}(f_i) = \operatorname{layer}(y_i(x)) = i.$$

We are now ready to prove Theorem 1.

Proof of Theorem 1. For $i < j$, let $\mathcal{P}$ be a path between any $y_i(x') \in y_i$ and $y_j(x) \in \mathcal{D}_{i+1:N}$. Let $y_k(\hat{x})$ be the first node in $\mathcal{P}$ such that $\operatorname{layer}(y_k(\hat{x})) > i$. Then $\mathcal{P}$ contains

$$\cdots \to y_m(\hat{x}) \to y_k(\hat{x}) \rightleftharpoons \cdots$$

for some $m \leq i < k$.

If $y_k(\hat{x}) \in \mathcal{D}_{i+1:N}$, then, since $\mathcal{D}$ is closed downwards, $y_m(\hat{x}) \in \mathcal{D}_{1:i}$, meaning that $\mathcal{P}$ is inactive.

If, on the other hand, $y_k(\hat{x}) \notin \mathcal{D}$, then, since $\mathcal{D}$ is closed downwards, $y_{k'}(\hat{x}) \notin \mathcal{D}$ for all $k' \geq k$. Therefore, $y_j(x)$ cannot be descendant of $y_k(\hat{x})$, so $\mathcal{P}$ must contain

$$\cdots \to y_{k'}(\hat{x}) \to y_{k''}(\hat{x}) \leftarrow f_{k''} \to \cdots$$

for some $m \leq k' < k''$, which forms a v-structure. We conclude that $\mathcal{P}$ is inactive, because $y_{k''}(\hat{x})$ nor a descendant of $y_{k''}(\hat{x})$ can be in $\mathcal{D}$. □

B The Nonlinear and Linear Equivalent Model

In this section, we construct equivalent models for GPAR (Lemmas 1 and 2). These models make GPAR's connection to other models in the literature explicit.

To begin with, we must introduce some notation and definitions. For functions $A, B \colon \mathcal{X} \times (\mathcal{Y}^M)^{\mathcal{X}} \to \mathcal{Y}^M$, define composition $\circ$ as follows: $(A \circ B)(x, y) = A(x, B(\cdot, y))$. Note that $\circ$ is well-defined and associative. For a function $u \colon \mathcal{X} \to \mathcal{Y}^M$, denote $A \circ u \colon \mathcal{X} \to \mathcal{Y}^M$, $A \circ u = A(\cdot, u)$. Again, note that $(A \circ B) \circ u = A \circ (B \circ u)$. Furthermore, denote

$$\underbrace{A \circ \cdots \circ A}_{n \text{ times}} = A^n.$$

Consider a function $A\colon \mathcal{X} \times (\mathcal{Y}^M)^{\mathcal{X}} \to \mathcal{Y}^M$ such that $A_i(x, y)\colon \mathcal{X} \times (\mathcal{Y}^M)^{\mathcal{X}} \to \mathcal{Y}$ depends only on $(x, y_{1:i-1})$, where $A_1 = 0$. Further let $u, y\colon \mathcal{X} \to \mathcal{Y}^M$, denote $\mathrm{T}\, f = u + A \circ f$, and denote N consecutive applications of T by T^N.

The expression $\mathrm{T}^{M-1}\, u$ will be key in constructing the equivalent models. We show that it is the unique solution of a functional equation:

Proposition 1. The unique solution of $y = u + A \circ y$ is $y = \mathrm{T}^{M-1}\, u$.

Proof of Proposition 1. First, we show that $y = u + A \circ y$ has a solution, and that this solution is unique. Because $A_i(x, y)$ depends only on $(x, y_{1:i-1})$, it holds that

$$y_i = u_i + A_i \circ y = u_i + A_i \circ (y_{1:i-1}, 0),$$

where $(y_{1:i-1}, 0)$ represents the concatenation of $y_{1:i-1}$ and $M - i + 1$ zeros. Thus, y_i can uniquely be constructed from u_i, A_i, and $y_{1:i-1}$; therefore, y_1 exists and is unique, so y_2 exists and is unique: by induction we find that y exists and is unique.

Second, we show that $y = \mathrm{T}^{M-1}\, u$ satisfies $y = u + A \circ y = \mathrm{T}\, y$. To show this, we show that $(\mathrm{T}^n\, u)_i = (\mathrm{T}^{n-1}\, u)_i$ for $i = 1, \ldots, n$, for all n. To begin with, we show the base case, $n = 1$:

$$(\mathrm{T}\, u)_1 = u_1 + A_1 \circ u = u_1 = (\mathrm{T}^0\, u)_1,$$

since $A_1 = 0$. Finally, suppose that the claim holds for a particular n. We show that the claim then holds for $n + 1$: Let $i \leq n + 1$. Then

$$\begin{aligned}
(\mathrm{T}^{n+1}\, u)_i &= u_i + A_i \circ \mathrm{T}^n\, u \\
&= u_i + A_i \circ ((\mathrm{T}^n\, u)_{1:i-1}, (\mathrm{T}^n\, u)_{i:M}) \\
&= u_i + A_i \circ ((\mathrm{T}^{n-1}\, u)_{1:i-1}, (\mathrm{T}^n\, u)_{i:M}) \\
&\qquad\qquad \text{(By assumption)} \\
&\overset{(*)}{=} u_i + A_i \circ ((\mathrm{T}^{n-1}\, u)_{1:i-1}, (\mathrm{T}^{n-1}\, u)_{i:M}) \\
&= u_i + A_i \circ \mathrm{T}^{n-1}\, u \\
&= (\mathrm{T}^n\, u)_i,
\end{aligned}$$

where $(*)$ holds because $A_i(x, y)$ depends only on $(x, y_{1:i-1})$. □

In the linear case, $\mathrm{T}^{M-1}\, u$ turns out to greatly simplify.

Proposition 2. If $A(x, y)$ is linear in y, then $\mathrm{T}^{M-1}\, u = (\sum_{i=0}^{M-1} A^i) \circ u$.

Proof of Proposition 2. If $A(x, y)$ is linear in y, then one verifies that $\circ$ distributes over addition. Therefore,

$$\begin{aligned}
\mathrm{T}^{M-1}\, u &= u + A \circ \mathrm{T}^{M-2}\, u \\
&= u + A \circ u + A^2 \circ \mathrm{T}^{M-3}\, u \\
&\;\;\vdots \\
&= u + A \circ u + \cdots + A^{M-1} \circ u. \qquad \square
\end{aligned}$$

We now use Propositions 1 and 2 to construct a nonlinear and linear equivalent model.

Lemma 1 (Nonlinear Equivalent Model)**.** Let A be an M-dimensional vector-valued process over $\mathcal{X} \times (\mathcal{Y}^M)^{\mathcal{X}}$, each A_i drawn from $\mathcal{GP}(0, k_{A_i})$ independently, and let u be an M-dimensional vector-valued process over $\mathcal{X}$, each u_i drawn from $\mathcal{GP}(0, k_{u_i})$ independently. Furthermore, let $A_i(x, y)\colon \mathcal{X} \times (\mathcal{Y}^M)^{\mathcal{X}} \to \mathcal{Y}$ depend only on $(x, y_{1:i-1})$, meaning that $k_{A_i} = k_{A_i}(x, y_{1:i-1}, x', y'_{1:i-1})$, and let $A_1 = 0$. Denote $\mathrm{T}\, f = u + A \circ f$, and denote N consecutive applications of T by T^N. Then

$$\begin{aligned}
y \,|\, A, u = \mathrm{T}^{M-1}\, u \iff \\
y_i \,|\, y_{1:i-1} \sim \mathcal{GP}(0, k_{u_i} + k_{A_i}(\cdot\,, y_{1:i-1}, \cdot\,, y_{1:i-1})).
\end{aligned}$$

Proof of Lemma 1. Since $A_i(x, y)$ depends only on $(x, y_{1:i-1})$, it holds by Proposition 1 that any sample from $y \,|\, A, u$ satisfies $y_i = u_i + A_i \circ y$, so $y_i = u_i + A_i \circ (y_{1:i-1}, 0)$, where $(y_{1:i-1}, 0)$ represents the concatenation of $y_{1:i-1}$ and $M - i + 1$ zeros. The equivalence now follows. □

Lemma 2 (Linear Equivalent Model)**.** Suppose that A was instead generated from

$$A(x, y) \,|\, \hat{A} = \int \hat{A}(x - z) y(z)\, \mathrm{d}z,$$

where $\hat{A}$ is an $(M \times M)$-matrix-valued process over $\mathcal{X}$, each $\hat{A}_{i,j}$ drawn from $\mathcal{GP}(0, k_{\hat{A}_{i,j}})$ independently if $i > j$ and $\hat{A}_{i,j} = 0$ otherwise. Then

$$\begin{aligned}
y \,|\, A, u = \left(\sum_{i=0}^{M-1} A^i \right) \circ u \iff \\
y_i \,|\, y_{1:i-1} \sim \mathcal{GP}(0, k_{u_i} + k_{A_i}(\cdot\,, y_{1:i-1}, \cdot\,, y_{1:i-1})), \quad (1)
\end{aligned}$$

where

$$\begin{aligned}
&k_{A_i}(x, y_{1:i-1}, x', y'_{1:i-1}) \\
&\quad = \sum_{j=1}^{i-1} \int k_{\hat{A}_{i,j}}(x - z, x' - z') y_j(z) y'_j(z')\, \mathrm{d}z\, \mathrm{d}z'.
\end{aligned}$$

Proof of Lemma 2. First, one verifies that $A_i(x, y)$ still depends only on $(x, y_{1:i-1})$, and that $A_i(x, y)$ is linear in y. The result then follows from Lemma 1 and Proposition 2, where the expression for k_{A_i} follows from straightfoward calculation. □

As mentioned in the paper, the kernels for $f_{1:M}$ determine the types of relationships between inputs and outputs that can be learned. Lemmas 1 and 2 make this explicit: Lemma 1 shows that nonlocal GPAR can recover a model where M latent GPs u are repeatedly composed with another latent GP A, where A has a particular dependency structure, and Lemma 2 shows that nonlocal GPAR can recover a model where M latent GPs u are linearly transformed, where the linear transform $T = \sum_{i=0}^{M-1} A^i$ is lower triangular and may vary with the input.

In Lemma 2, note that it is not restrictive that T is lower triangular: Suppose that T were dense. Then, letting $y \,|\, T, u = T \circ u$, $y \,|\, T$ is jointly Gaussian. Hence $y_i \,|\, y_{1:i-1}, T$ is a GP whose mean linearly depends upon $y_{1:i-1}$ via T, meaning that $y_i \,|\, y_{1:i-1}$ is of the form of Equation (1) where k_{u_i} may be more complicated. If, however, $\hat{A}(z) = \delta(z)B$ for some random $(M \times M)$-matrix B, each $B_{i,j}$ drawn from $\mathcal{N}(0, \sigma^2_{B_{i,j}})$ if $i > j$ and $B_{i,j} = 0$ otherwise, then it is restrictive that T is lower triangular: In this case, $y(x) \,|\, B, u = \sum_{i=0}^{M-1} B^i u(x)$. If $T = \sum_{i=0}^{M-1} B^i$ were dense, then, letting $y \,|\, T, u = Tu$, y can be represented with Lemma 2 if and only if $y \,|\, T$'s covariance can be diagonalised by a constant, invertible, lower-triangular matrix. This condition does not hold in general, as Lemma 3 proves.

C Lemma 3

Call functions $k_1, \ldots, k_M \colon \mathcal{X} \to \mathbb{R}$ linearly independent if

$$\left(\forall\, x : \sum_{i=1}^{M} c_i k_i(x) = 0 \right) \implies c_1 = \ldots = c_M = 0.$$

Lemma 3. Let $k_1, \ldots, k_M \colon \mathcal{X} \to \mathbb{R}$ be linearly independent and arrange them in a diagonal matrix $K = \operatorname{diag}(k_1, \ldots, k_n)$. Let A be an invertible $M \times M$ matrix such that its columns cannot be permuted into a triangular matrix. Then there does not exist an invertible triangular matrix T such that $T^{-1} B K(x) B^\mathsf{T} T^{-\mathsf{T}}$ is diagonal for all x.

Proof. Suppose, on the contrary, that such T does exist. Then two different rows a_p and a_q of $A = T^{-1}B$ share nonzero elements in some columns C; otherwise, A would have exactly one nonzero entry in every column—A is invertible—so A would be the product of a permutation matrix and a diagonal matrix, meaning that $B = TA$'s columns could be permuted into a triangular matrix. Now, by $T^{-1} B K(x) B^\mathsf{T} T^{-\mathsf{T}} = A K(x) A^\mathsf{T}$ being diagonal for all $x \in \mathcal{X}$, $\sum_i a_{p,i} a_{q,i} k_i(x) = 0$ for all x. Therefore, by linear independence of $k_1, \ldots, k_N$, it holds that $a_{p,i} a_{q,i} = 0$ for all i. But $a_{p,i} a_{q,i} \neq 0$ for any $i \in C$, which is a contradiction. □

D Experimental Details

For every experiment, the form of the kernels is determined by the particular GPAR model used: GPAR-L, GPAR-NL, or GPAR-L-NL (see Table 2 in the main paper), potentially with a D-$*$ prefix to indicate the denoising procedure outlined in "Potential deficiencies of GPAR" in Section 2 of the paper main. For GPAR-NL, we always used exponentiated quadratic (EQ) kernels, except for the exchange rates experiment, where we used rational quadratic (RQ) kernels (Rasmussen and Williams, 2006). Furthermore, in every problem we simply expanded according to Equation (1) in the main paper or greedily optimised the ordering, in both cases putting the to-be-predicted outputs last. We used `scipy`'s implementation of the L-BFGS-B algorithm (Nocedal and Wright, 2006) to optimise hyperparameters.

Acknowledgements

Richard E. Turner is supported by Google as well as EPSRC grants EP/M0269571 and EP/L000776/1.

References

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, second edition, 2006.