Kernel Matrices

Define a kernel $\boldsymbol{\kappa} = \langle \kappa_1, \kappa_2, \ldots, \kappa_j \rangle$ as a vector satisfying $\kappa_i \geq 0, \forall i \in [0, j]$ and $\sum_{i=0}^{j} \kappa_i = 1$, where $j = |\boldsymbol{\kappa}|$ is the length of the kernel.

Given kernels $\mathcal{K} = \{\boldsymbol{\kappa}^{(1)}, \boldsymbol{\kappa}^{(2)}, \cdots, \boldsymbol{\kappa}^{(m)}\}$, set $\ell = \max\{|\kappa_i|, i \in [1, m]\}$ and define the zero-padded kernels copies

$$\tilde{\boldsymbol{\kappa}}^{(i)} = \begin{bmatrix} \boldsymbol{\kappa}^{(i)} \\ \vec{\mathbf{0}}_{\ell - |\boldsymbol{\kappa}^{(i)}|} \end{bmatrix} \in \mathbb{R}^{\ell}, \quad i = 1, \ldots, m.$$

Where $\vec{\mathbf{0}}_n$ is a zero vector of length n. Consequently, $\forall i, j \in [1, m], |\tilde{\boldsymbol{\kappa}}^{(i)}| = |\tilde{\boldsymbol{\kappa}}^{(j)}|$.

Stacking these transformed vectors yields the *kernel matrix*

$$\mathbf{K} = \begin{bmatrix} \tilde{\boldsymbol{\kappa}}^{(1)} & \tilde{\boldsymbol{\kappa}}^{(2)} & \ldots & \tilde{\boldsymbol{\kappa}}^{(m)} \end{bmatrix} \in \mathbb{R}^{\ell \times m}$$

Whose i-th column is a kernel.

Paramaterized Kernels

In a GKR model, each kernel $\tilde{\boldsymbol{\kappa}}^{(i)}$ is paramaterized by a discrete distribution or a continuous which is discretized to integer lags. Our software currently supports discrete Gaussian, Triangular and Gamma-distributed kernels. Consequently, each kernel $\tilde{\boldsymbol{\kappa}}^{(i)}$ is parameterized by a set of one or more parameters $\boldsymbol{\rho}^{(i)}$. For instance, should the kernel be Gamma-distributed with shape a and rate λ, one may let $\boldsymbol{\rho}^{(i)} = \langle a^{(i)}, \lambda^{(i)} \rangle$, the ℓ-th element $\kappa_\ell^{(i)} \in \boldsymbol{\kappa}^{(i)}$ would be defined as

$$\kappa_\ell^{(i)} = \int_i^{i+1} \frac{(\lambda^{(i)})^{a^{(i)}}}{\Gamma(a^{(i)})} x^{a^{(i)}-1} e^{-\lambda^{(i)} x} \, \mathrm{d}x = \frac{1}{\Gamma(a^{(i)})} \Big(\Gamma(a^{(i)}, \lambda^{(i)} \ell) - \Gamma(a^{(i)}, \lambda^{(i)}(\ell+1)) \Big)$$

Where $\Gamma(b, c)$ is the lower incomplete gamma function. Alternatively, should one desire a distribution-free kernel, the length k can be declared as a hyper-parameter, then $\boldsymbol{\rho}^{(i)} = \langle \rho_1, \rho_2, \ldots, \rho_k \rangle$ where each $\rho \in \boldsymbol{\rho}$ directly represents the kernel weights and are optimized independently.

Regardless of the chosen distributions, the kernels are normalized to sum to 1 after evaluation. The list of all kernel parameters must be optimized alongside the regression coefficients. Denote the list of kernel parameters as $\mathcal{P} = \left\{ \boldsymbol{\rho}^{(1)}, \boldsymbol{\rho}^{(2)}, \ldots, \boldsymbol{\rho}^{(k)} \right\}$ with length at least k.

Covariate-Wise Convolution

Define $\mathbf{K}_t$ as the previously defined $\mathbf{K}$, except the padding length ℓ is fixed at time $t \geq \ell$. Similarly, let $\mathbf{X}_t \subseteq \mathbf{X}$ be the data matrix containing observations up to and including time t.

Consequently, we define

$$(\mathbf{X} * \mathbf{K})[t] = \mathbf{1}_t^\top \big((\mathbf{P}\mathbf{X}_t) \circ \mathbf{K}_t\big) \tag{1}$$

Where $\mathbf{A} \circ \mathbf{B}$ is the Hadamard product of $\mathbf{A}$ and $\mathbf{B}$, and $\mathbf{P}$ is the $t \times t$ anti-identity matrix defined by

$$p_{ij} = \begin{cases} 1, & i + j = t + 1 \\ 0, & \text{otherwise} \end{cases}$$

By construction, any fixed column i of $(\mathbf{X} * \mathbf{K})[t]$ is the discrete convolution of the i-th covariate of $\mathbf{X}$ with the i-th kernel in $\mathbf{K}$ at time t.

GKR Algorithm as a Modified Data Matrix

Let $\mathbf{X}_t = x_i^{(j)}$, $i = 1, 2, \ldots t$, $j = 1, 2, \ldots k$ be a data matrix containing observations up to and including time t. Let $\mathbf{K}_t = \begin{bmatrix} \tilde{\boldsymbol{\kappa}}^{(1)} & \ldots & \tilde{\boldsymbol{\kappa}}^{(k)} \end{bmatrix}$ be a kernel matrix with each entry padded to length t, as previously defined.

Consider the following expansion of Equation 1,

$$\begin{aligned}
(\mathbf{X} * \mathbf{K})[t] &= \left(\mathbf{1}_t^\top \big((\mathbf{P}\mathbf{X}_t) \circ \mathbf{K}_t\big)\right) \\
&= \left(\mathbf{1}_t^\top \left(\left(\begin{bmatrix} 0 & 0 & \ldots & 0 & 1 \\ 0 & 0 & \ldots & 1 & 0 \\ \vdots & \vdots & \iddots & \vdots & \vdots \\ 1 & 0 & 0 & \ldots & 0 \end{bmatrix} \begin{bmatrix} x_0^{(1)} & x_0^{(2)} & \ldots & x_0^{(k)} \\ x_1^{(1)} & x_1^{(2)} & \ldots & x_1^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ x_t^{(1)} & x_t^{(2)} & \ldots & x_t^{(k)} \end{bmatrix}\right) \circ \begin{bmatrix} \kappa_0^{(1)} & \kappa_0^{(2)} & \ldots & \kappa_0^{(k)} \\ \kappa_1^{(1)} & \kappa_1^{(2)} & \ldots & \kappa_1^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ \kappa_t^{(1)} & \kappa_t^{(2)} & \ldots & \kappa_t^{(k)} \end{bmatrix}\right)\right) \\
&= \left(\begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} x_t^{(1)}\kappa_0^{(1)} & x_t^{(2)}\kappa_0^{(2)} & \ldots & x_t^{(k)}\kappa_0^{(k)} \\ x_{t-1}^{(1)}\kappa_1^{(1)} & x_{t-1}^{(2)}\kappa_1^{(2)} & \ldots & x_{t-1}^{(k)}\kappa_1^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ x_0^{(1)}\kappa_t^{(1)} & x_0^{(2)}\kappa_t^{(2)} & \ldots & x_0^{(k)}\kappa_t^{(k)} \end{bmatrix}\right) \\
&= \begin{bmatrix} \sum_{i=0}^t x_{t-i}\kappa_i^{(1)} & \sum_{i=0}^t x_{t-i}\kappa_i^{(1)} & \ldots & \sum_{i=0}^t x_{t-i}\kappa_i^{(k)} \end{bmatrix}
\end{aligned}$$

Since each element of the data matrix is some real number, given the kernel parameters $\mathcal{P}$ the above $(\mathbf{X} * \mathbf{K})[t]$ evaluates to a row vector of length k.

With this in mind, consider the block matrix defined by

$$\mathbf{S} = \begin{bmatrix} (\mathbf{X} * \mathbf{K})[1] \\ (\mathbf{X} * \mathbf{K})[2] \\ \vdots \\ (\mathbf{X} * \mathbf{K})[T] \end{bmatrix}$$

Where T is the length of the response and covariate time series. Noting the simplification of $(\mathbf{X} * \mathbf{K})[t]$ it follows that $\mathbf{S}$ is of the form

$$\mathbf{S} = \begin{bmatrix} \sum_{i=0}^{1} x_{1-i}\kappa_i^{(1)} & \sum_{i=0}^{1} x_{1-i}\kappa_i^{(1)} & \dots & \sum_{i=0}^{1} x_{1-i}\kappa_i^{(k)} \\ \sum_{i=0}^{2} x_{2-i}\kappa_i^{(1)} & \sum_{i=0}^{2} x_{2-i}\kappa_i^{(1)} & \dots & \sum_{i=0}^{2} x_{2-i}\kappa_i^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=0}^{T} x_{T-i}\kappa_i^{(1)} & \sum_{i=0}^{T} x_{T-i}\kappa_i^{(1)} & \dots & \sum_{i=0}^{T} x_{T-i}\kappa_i^{(k)} \end{bmatrix}$$

Consequently, given kernel parameters $\mathcal{P}$, the matrix of convolutions $\mathbf{S}$ forms a variant of a data matrix.

Toy Example

Suppose that we have two time series, each of which are of length $T = 3$. In this simple setting, we allow $\mathbf{X}$ and $\mathbf{K}$ to be defined as

$$\mathbf{X} = \begin{bmatrix} 2 & 1 \\ 3 & -1 \\ 1 & 5 \end{bmatrix}, \qquad \mathbf{K} = \begin{bmatrix} 0.7 & 0.2 \\ 0.3 & 0.8 \end{bmatrix}$$

The computation of the first row of $\mathbf{S}$ is trivial. More interesting are the second and third rows. Consider the second row of $\mathbf{S}$.

$$\mathbf{S}_2 = \mathbf{1}_2^\top \left(\begin{bmatrix} 3 & -1 \\ 2 & 1 \end{bmatrix} \circ \begin{bmatrix} 0.7 & 0.2 \\ 0.3 & 0.8 \end{bmatrix} \right) = \begin{bmatrix} 2.7 & 0.6 \end{bmatrix}$$

The zero-padded kernels within $\mathbf{K}$ are implicit by definition of $\mathbf{K}_3$. Consequently, the third row of $\mathbf{S}$ is

$$\mathbf{S}_3 = \mathbf{1}_3^\top \left(\begin{bmatrix} 1 & 5 \\ 3 & -1 \\ 2 & 1 \end{bmatrix} \circ \begin{bmatrix} 0.7 & 0.2 \\ 0.3 & 0.8 \\ 0 & 0 \end{bmatrix} \right) = \begin{bmatrix} 1.6 & 0.2 \end{bmatrix}$$

Thus, the adjusted data matrix in this case is

$$\mathbf{S} = \begin{bmatrix} 1.4 & 0.2 \\ 2.7 & 0.6 \\ 1.6 & 0.2 \end{bmatrix}$$

This is used as a test case for our computation function.

Fitting a GKR Model with ARMA Errors: Gamma Case

A Generalized Linear Autoregressive Moving Average (GLARMA) model is defined by the linear predictor W_t, which in a GKR context is given by

$$W_t \mid \mathcal{F}_t = \mathbf{s}_t\boldsymbol{\beta} + Z_t, \text{ where } Z_t = \sum_{i=1}^{p} \phi_i(Z_{t-i} + e_{t-i}) + \sum_{j=1}^{q} \theta_j e_{t-j}$$

Where e_t is the Pearson error at time t. The authors require the response $\mathbf{y}_T$ to take an exponential-family distribution. Specifically, one of the modified form

$$f_Y(y \mid W_t) = \exp\{y_t W_t - a_t b(W_t) + c_t\}$$

Where, for a Gamma-distributed random variable, $a_t = \alpha$, $b(W) = -\log(-W)$ (canonical inverse), and $c(\alpha, y_t) = (\alpha - 1)\log(y) - \log(\Gamma(\alpha)) - \alpha\log(\alpha)$.

Consequently, the mean relationship of the predictors with the response would be

$$\mu_t = -\frac{1}{\mathbf{s}_t^\top \boldsymbol{\beta} + Z_t}$$

Hence, each beta term has a reciprocal relationship with μ. Because the canonical parameter is negative when $\mu > 0$, the sign of β has the opposite effect on μ compared to the canonical parameter.

However, should we wish to use these links, a modification of Equations 14-19 would allow us to estimate GLARMA models in our construction.