Estimating a Gaussian precision kernel with covariate information

Wicher Bergsma

October 3, 2023

Abstract

Assuming the precision kernel to be in a conditional RKKS, we propose a methodology to estimate it and its hyperparameters. The present work can be viewed as an extension of the I-prior methodology for regression (Bergsma, 2019; Bergsma and Jamil, 2023) to covariance estimation. Applications are not only in time series analysis (when time is the covariate), but also in regression when we want a flexible estimation of the error covariance.

1 Models for positive definite kernels

Let $\mathcal{X}$ be a set and $\mathcal{F}$ a vector space of symmetric functions on $\mathcal{X} \times \mathcal{X}$, which we will refer to as kernels. We refer to such $\mathcal{F}$ as kernel spaces. The class of positive definite kernels in a kernel space $\mathcal{F}$, that is functions $f \in \mathcal{F}$ such that $\sum \alpha_i \alpha_j f(x_i, x_j) \geq 0$, forms a convex cone in $\mathcal{X}$.

We will consider three classes of kernel spaces and equip these with an inner product to form an RKKS: tensor product of RKHSs, RKKSs of stationary kernels, and a generalization of the latter, an RKKS of all positive definite kernels on a set $\mathcal{X}$.

Further classes of kernels are given by Genton (2001).

1.1 Reproducing kernel Krein spaces

1.2 Tensor product models

Let $\mathcal{X}$ be a set and let $\mathcal{F}$ be an RKKS on $\mathcal{X}$ with r.k. h. Then the tensor product $\mathcal{F} \otimes \mathcal{F}$ is an RKHS on $\mathcal{X} \times \mathcal{X}$ with reproducing kernel $h \otimes h$. Any such RKHS contains a convex cone of positive definite kernels, namely the closure of the positively weighted span of the $h(x_i, \cdot) \otimes h(x_i, \cdot)$. Note $\Theta(x, x') = \langle \Theta, h(x, \cdot) \otimes h(x', \cdot) \rangle_{\mathcal{F} \otimes \mathcal{F}}$.

1.3 Stationary kernels

We construct an RKKS which contains a convex cone of the set of symmetric positive definite kernels.

A function f on $\mathbb{R}^m$ is called positive definite if $f(x - x')$ is a positive definite kernel. Bochner's theorem states that any continuous positive definite function is the Fourier transform of a positive measure:

$$f(t) = \int e^{i\,t \cdot u} \mu_f(u) du$$

We have via the inverse Fourier transform

$$\mu_f(t) = \int e^{-it\cdot u} f(u) du$$

An RKKS which contains a convex cone consisting of the set of positive definite functions as follows.

Lemma 1. *The kernel h defined by $h(x,t) = e^{ix\cdot t}$ is the unique reproducing kernel of the RKKS with indefinite inner product*

$$\langle f, f' \rangle = \int f(x) f'(t) e^{-ix\cdot t} dx dt$$

The RKKS consists of functions possessing a Fourier transform.

Proof. Let δ_x be the delta function centred at x, which is in the RKKS as it has a Fourier transform. The function ϕ_x defined by $\phi_x(t) = e^{ix\cdot t}$ has a Fourier transform, $\sqrt{2\pi}\delta_x$, so is in the RKKS. Since

$$f(x) = \int e^{ix\cdot t} \mu_f(u) du = \int \int e^{ix\cdot t} e^{-it\cdot u} f(u) du dt = \langle f, \phi_x \rangle$$

the reproducing property is satisfied. □

Bochner's theorem then immediately yields an RKKS which contains the stationary kernels.

Corollary 1. *The kernel $h : (\mathcal{X} \times \mathcal{X})^2 \to \mathbb{C}$ defined by $h((x,t),(x',t')) = e^{i(x-t)\cdot(x'-t')}$ is the unique reproducing kernel of the RKKS with indefinite inner product*

$$\langle k, k' \rangle = \int k(x,t) k'(x',t') e^{-i(x-t)\cdot(x'-t')} dx dt dx' dt'$$

The RKKS contains all symmetric stationary kernels.

A stochastic process is called stationary if its covariance kernel is of the form $f(x - x')$ for a positive definite function f. The measure μ_f is then called the *spectral measure* of the process. Hence, we have a model which includes precisely the stationary Gaussian processes. Using the I-prior methodology of this paper, a single observed time series can then suffice to estimate both the stationary covariance kernel *and* the trend (with the trend estimated by the I-prior methodology of Bergsma (2019); Bergsma and Jamil (2023)).

1.4 Model for nonstationary kernels based on spectral representation

A kernel is positive definite if and only if it has the form (Yaglom, 1987)

$$h(x,t) = \int \int e^{i(\omega_1^\top x - \omega_2^\top t)} \mu(d\omega_1, d\omega_2)$$

for a nonnegative symmetric measure μ. These are a convex cone of the RKKS with kernel $g((x,t),(x',t')) = e^{i(x\cdot x' - t\cdot t')}$?

2 Likelihood

Let $\mathcal{X}$ be a set and let $\Theta : \mathcal{X} \times \mathcal{X}$ be a symmetric and positive definite kernel, i.e., $\Theta(x, x') = \Theta(x', x)$ and $\sum_{t,u=1}^{n} \alpha_t \alpha_u \Theta(x_t, x_u) \geq 0$ for all $\alpha_t \in \mathbb{R}$, $x_t \in \mathcal{X}$, $n = 1, 2, \ldots$. For $x = (x_1, \ldots, x_m) \in \mathcal{X}^m$, a normal density for $(y|x) \in \mathbb{R}^m$ is given as

$$p(y|x, \Theta) = (2\pi)^{-m/2} \, |\Theta_x|^{1/2} \, e^{-\frac{1}{2} y^\top \Theta_x y}$$

where Θ_x is the $m \times m$ precision matrix with (t, u)th element $\Theta(x_t, x_u)$. The log-likelihood is

$$\ell(\Theta|x, y) = -\frac{m}{2} \log(2\pi) + \frac{1}{2} \log|\Theta_x| - \frac{1}{2} y^\top \Theta_x y$$

To be able to compute the score and the Fisher information for Θ, we need to make assumptions on the set of possible values Θ can take. A flexible class of sets is formed by RKKSs. Let $\mathcal{F}$ be an RKKS on $\mathcal{X} \times \mathcal{X}$ with reproducing kernel $h : (\mathcal{X} \times \mathcal{X})^2 \to \mathbb{C}$. Then note that $\Theta(x, x') = \langle \Theta, h(x, \cdot) \otimes h(x', \cdot) \rangle_{\mathcal{F}}$. Without loss of generality, we may assume symmetry in the arguments, i.e., $h((x, x'), (x'', x''')) = h((x', x), (x'', x'''))$.

The score function $s : \mathcal{F} \to \mathcal{F}$ then is given by

$$s(\Theta|x, y) = \sum_{t,u=1}^{m} (y_t y_u - \sigma_{t,u}) \, h((x_t, x_u), (\cdot, \cdot))$$

where $\sigma_{t,u} = E y_t y_u = \operatorname{cov}(y_t, y_u)$. If Θ_x is invertible, $\sigma_{t,u}$ is the (t, u)th element of its inverse. Recall that

$$\operatorname{cov}(y_t y_u, y_v y_r) = \sigma_{t,v} \sigma_{u,r} + \sigma_{t,r} \sigma_{u,v}.$$

Hence, the Fisher information on Θ, which is the covariance kernel of $s(\Theta)$, is given as

$$\begin{aligned}
\mathcal{I}(\Theta) &= E_{y \sim p} \big[s(\Theta) \otimes s(\Theta) \big] \\
&= \sum_{t,u} \sum_{v,r} (\sigma_{t,v} \sigma_{u,r} + \sigma_{t,r} \sigma_{u,v}) \, (h((x_t, x_u), (\cdot, \cdot)) \otimes h((x_v, x_r), (\cdot, \cdot)))
\end{aligned}$$

Here, $\mathcal{I}(\Theta)$ is understood as an element of $\mathcal{F} \otimes \mathcal{F}$, such that

$$\mathcal{I}(\Theta)((x, x'), (x'', x''')) = \sum_{t,u} \sum_{v,r} (\sigma_{t,v} \sigma_{u,r} + \sigma_{t,r} \sigma_{u,v}) \, (h((x_t, x_u), (x, x')) \otimes h((x_v, x_r), (x'', x''')))$$

2.1 Special cases

For the tensor product model, the score is

$$s(\Theta) = \sum_{t,u=1}^{m} (y_t y_u - \sigma_{t,u}) \, h(x_t, \cdot) \otimes h(x_u, \cdot)$$

where $\sigma_{t,u} = E y_t y_u = \operatorname{cov}(y_t, y_u)$ is the (t, u)th element of Θ_x^{-1}. Recall that

$$\operatorname{cov}(y_t y_u, y_v y_r) = \sigma_{t,v} \sigma_{u,r} + \sigma_{t,r} \sigma_{u,v}.$$

Hence, the Fisher information on Θ, which is the covariance kernel of $s(\Theta)$, is given as

$$\begin{aligned}
\mathcal{I}(\Theta) &= E_{y \sim p} \big[s(\Theta) \otimes s(\Theta) \big] \\
&= \sum_{t,u} \sum_{v,r} (\sigma_{t,v} \sigma_{u,r} + \sigma_{t,r} \sigma_{u,v}) \, (h(x_t, \cdot) \otimes h(x_u, \cdot)) \otimes (h(x_v, \cdot) \otimes h(x_r, \cdot))
\end{aligned}$$

Here, $\mathcal{I}(\Theta)$ is understood as an element of $(\mathcal{F} \otimes \mathcal{F}) \otimes (\mathcal{F} \otimes \mathcal{F})$, such that

$$\mathcal{I}(\Theta)((x, x'), (x'', x''')) = \sum_{t,u} \sum_{v,r} (\sigma_{t,v}\sigma_{u,r} + \sigma_{t,r}\sigma_{u,v}) h(x_t, x) h(x_u, x') h(x_v, x'') h(x_r, x''')$$

3 Models and hypothesis tests

3.1 Simple hypotheses

We may wish to test a hypothesis

$$H_0 : \Theta = \Theta_0$$

against

$$H_1 : \Theta \in \mathcal{F}_h$$

where $\mathcal{F}_h$ is the RKKS with r.k. h. The modified score test statistic is the *RKHS* norm of the score vector, and reduces to

$$T^2 = (S_{\mathbf{y}|x} - \Sigma_0)^\top H_x^* (S_{\mathbf{y}|x} - \Sigma_0)$$

where $\Sigma_0 = E_{H_0}(YY^\top)$ and $H_x^* = \sum_{i=1}^{n^2} |\lambda_i| u_i u_i^\top$ if $H_x = \sum_{i=1}^{n^2} \lambda_i u_i u_i^\top$. (Note these are $n^2 \times n^2$ matrices.)

More generally, with g an r.k. such that $\mathcal{F}_g \subset \mathcal{F}_h$, we can test

$$H_0 : \Theta \in \mathcal{F}_g$$

giving

$$T^2 = (S_{\mathbf{y}|x} - \hat{\Sigma}_0)^\top H_x^* (S_{\mathbf{y}|x} - \hat{\Sigma}_0)$$

where $\hat{\Sigma}_0$ is a suitable estimator under H_0.

We can test eg the following:

- Stationarity: take $g((x,t),(x',t')) = e^{i(x-t)(x'-t')}$ and $h(x,x') = e^{x \cdot x' - t \cdot t'}$.
- White noise: set $\Sigma_0 = I$ and h some appropriate kernel.
- A two sample test (whether two samples of processes have the same covariance kernel) is based on

  $$T^2 = \operatorname{tr}\left[(S_{\mathbf{y}_1|x} - S_{\mathbf{y}_2|x}) H_x (S_{\mathbf{y}_1|x} - S_{\mathbf{y}_2|x}) H_x\right]$$

  A permutation significance test can be done.

These tests differ from existing ones in that they take into account the kernel.

Question: test using bootstrap? Other method?

3.2 Models

We can assume $\mathcal{F}$ is an interaction space on $\mathcal{X}$, for example of the form

$$\mathcal{F} = \mathcal{C}_1 \otimes \mathcal{C}_2 + \mathcal{C}_1 \otimes \mathcal{F}_2 + \mathcal{F}_1 \otimes \mathcal{C}_2$$

or

$$\mathcal{F} = \mathcal{C}_1 \otimes \mathcal{C}_2 + \mathcal{C}_1 \otimes \mathcal{F}_2 + \mathcal{F}_1 \otimes \mathcal{C}_2 + \mathcal{F}_1 \otimes \mathcal{F}_2$$

We can test which model holds.

4 Dimension reduction

We show that, for estimation purposes, it suffices to estimate an $m \times m$ matrix of unknowns.

Let $x = (x_1, \ldots, x_m) \in \mathcal{X}^m$ and define

$$\mathcal{F}_m = \{f \in \mathcal{F} | f(x) = \sum_{t=1}^{m} f(x_t) w_t\}$$

Its orthogonal complement in $\mathcal{F}$ is

$$\mathcal{F}_m^{\perp} = \{f \in \mathcal{F} | f(x_t) = 0, t = 1, \ldots, m\}$$

The tensor product space $\mathcal{F}_m \otimes \mathcal{F}_m$ is finite dimensional:

$$\mathcal{F}_m \otimes \mathcal{F}_m = \Big\{ f \in \mathcal{F} \otimes \mathcal{F} \,\Big|\, f(x, x') = \sum_{t,u=1}^{m} w_{t,u}\, h(x_t, \cdot) \otimes h(x_u, \cdot) \Big\}$$

Its orthogonal complement in $\mathcal{F} \otimes \mathcal{F}$ is

$$\big(\mathcal{F}_m \otimes \mathcal{F}_m\big)^{\perp} = \Big\{ f \in \mathcal{F} \otimes \mathcal{F} \,\Big|\, f(x_t, x_u) = 0, t = 1, \ldots, m, u = 1, \ldots, m \Big\}$$

We can uniquely decompose Θ as

$$\Theta = \Theta_m + R_m, \quad \Theta_n \in \mathcal{F}_m \otimes \mathcal{F}_m, R_m \in \big(\mathcal{F}_m \otimes \mathcal{F}_m\big)^{\perp}$$

If Θ is a precision kernel,

$$\Theta_x = \sum_{t,u=1}^{m} w_{t,u} h(x_t, \cdot) \otimes h(x_u, \cdot)$$

where $W = (w_{t,u})$ is symmetric and positive definite

The likelihood depends on the $\Theta(x_t, x_u)$. However, $\Theta(x_t, x_u) = \Theta_m(x_t, x_u)$, i.e., the likelihood does not depend on R_m. Since additionally $R_m \perp \Theta_m$, there is no Fisher information on R_m, that is, without further prior information R_m cannot be estimated from the data, and we set it to 0.

5 I-priors

Assume observations $y_i | x$, $i = 1, \ldots, n$, where $y_i \in \mathbb{R}^m$ and denote $\mathbf{y} = (y_1, \ldots, y_n)$. Here, x is the covariate and is the same for each y_i. The aim is to estimate the precision kernel Θ assuming Gaussianity and taking the covariate information x into account. The log-likelihood becomes

$$\ell(\Theta | x, \mathbf{y}) = -\frac{m}{2} \log(2\pi) + \frac{1}{2} \log|\Theta_x| - \frac{1}{2} \operatorname{tr}(S_{\mathbf{y}|x} \Theta_x)$$

where $S_{\mathbf{y}|x} = n^{-1} \sum y_i y_i^{\top}$.

The Wishart distribution is a conjugate prior for a normal $m \times m$ precision matrix. We show it can be used for a precision kernel as well, so that the posterior can be used to make predictions at not previously observed covariate values.

As outlined above, to estimate the kernel $\Theta \in \mathcal{F} \otimes \mathcal{F}$, we only need to estimate the $m \times m$ symmetric positive definite matrix W. Let us assign a (conjugate) Wishart prior distribution $\text{Wish}(W_0, \nu)$ with density

$$\pi(W|W_0, \nu) = 2^{-\nu m/2} \Gamma(\nu/2) |W_0|^{-\nu/2} |W|^{\nu - m - 1)/2} e^{-\frac{1}{2} \operatorname{tr}(W_0^{-1} W)}$$

The posterior then is

$$\pi(W|x, \mathbf{y}, W_0, \nu) = \text{Wish}((nH_x S_{\mathbf{y}|x} H_x + W_0^{-1})^{-1}, n + \nu)$$

and the posterior mean is

$$\hat{W} = (n + \nu)(nH_x S_{\mathbf{y}|x} H_x + W_0^{-1})^{-1}$$

The posterior for Θ has an extended Wishart distribution,

$$\Theta|x, \mathbf{y} \sim \sum_{t,u=1}^{n} (w_{t,u}|x, \mathbf{y}) h(x_t, \cdot) \otimes h(x_u, \cdot)$$

where $W|x, \mathbf{y} = (w_{t,u})|x, \mathbf{y} \sim \text{Wish}((n + \nu)^{-1} \hat{W}, n + \nu)$. We may denote this as

$$\Theta|x, \mathbf{y} \sim \text{Wish}\left((n + \nu)^{-1} \sum_{t,u=1}^{n} \hat{w}_{t,u} h(x_t, \cdot) \otimes h(x_u, \cdot), n + \nu \right)$$

The question now is how to choose W_0. An I-prior is a prior for a parameter such that the covariance kernel of the parameter under the prior is proportional to its Fisher information under the model. The empirical I-prior sets $W_0 = S_{\mathbf{y}|x}$. This results in the prior covariance kernel for Θ equalling its empirical Fisher information. The I-prior can be interpreted as a maximum entropy prior... (how???), that is, the W_0 maximizing entropy...

6 Estimating hyperparameters

6.1 Marginal likelihood

The marginal likelihood is

$$\log p(W) = C + \frac{1}{2} m(n + \nu) \log(2) + \log[\Gamma((n + \nu)/2)] - \frac{n + \nu}{2} \log|\hat{W}| + g(\nu)$$

6.2 EM algorithm

The complete data log-likelihood is

$$\begin{aligned}
&\ell(\Theta|x, \mathbf{y}) + \log \pi(\Theta) \\
&\quad = C + \frac{1}{2} \log|W| - \frac{1}{2} \operatorname{tr}(S_{\mathbf{y}|x} \Theta_x) - \frac{1}{2} \nu m \log(2) + \log[\Gamma(\nu/2)] - \frac{\nu}{2} \log|S_{\mathbf{y}|x}| + \frac{1}{2}(\nu - m - 1) \log|W| - \frac{1}{2} \operatorname{tr}(S_{\mathbf{y}|x}^{-1} \\
&\quad = C + \frac{1}{2}(\nu - m - 2) \log|W| - \frac{1}{2} \operatorname{tr}([H_x S_{\mathbf{y}|x} H_x + S_{\mathbf{y}|x}^{-1}] W) - \frac{1}{2} \nu m \log(2) + \log[\Gamma(\nu/2)] - \frac{\nu}{2} \log|S_{\mathbf{y}|x}| \\
&\quad =: C + \frac{1}{2}(\nu - m - 2) \log|W| - \frac{1}{2} \operatorname{tr}([H_x S_{\mathbf{y}|x} H_x + S_{\mathbf{y}|x}^{-1}] W) + g(\nu)
\end{aligned}$$

where $\Theta_x = H_x W H_x$. We have $E(W|\mathbf{y}) = (n+\nu)(H_x S_{\mathbf{y}|x} H_x + S_{\mathbf{y}|x}^{-1})^{-1}$ and $E(\log|W|\,|\mathbf{y}) = \psi_m(\nu/2) + m\log(2) - \log|H_x S_{\mathbf{y}|x} H_x + S_{\mathbf{y}|x}^{-1}|$. Hence,

$$
\begin{aligned}
Q(\lambda) &= C + \frac{1}{2}(\nu - m - 2)(\psi_m(\nu/2) + m\log(2) - \log|H_x S_{\mathbf{y}|x} H_x + S_{\mathbf{y}|x}^{-1}|) \\
&\quad - \frac{n+\nu}{2}\operatorname{tr}([H_x S_{\mathbf{y}|x} H_x + S_{\mathbf{y}|x}^{-1}](H_x S_{\mathbf{y}|x} H_x + S_{\mathbf{y}|x}^{-1})^{-1}) + g(\nu) \\
&= C + \frac{1}{2}(\nu - m - 2)(\psi_m(\nu/2) + m\log(2) - \log|H_x S_{\mathbf{y}|x} H_x + S_{\mathbf{y}|x}^{-1}|) - \frac{m(n+\nu)}{2} + g(\nu) \\
&= C' + \frac{1}{2}(\nu - m - 2)(\psi_m(\nu/2) + m\log(2) - \log|H_x S_{\mathbf{y}|x} H_x + S_{\mathbf{y}|x}^{-1}|) - \frac{m\nu}{2} + g(\nu)
\end{aligned}
$$

Then

$$
\frac{dQ}{d\lambda} = -\frac{1}{2}\operatorname{tr}\left[(H_x S_{\mathbf{y}|x} H_x + S_{\mathbf{y}|x}^{-1})^{-1}\frac{d}{d\lambda} H_x S_{\mathbf{y}|x} H_x\right]
$$

which can be solved potentially efficiently.

7 Normal-Wishart priors

We may want to simultaneously estimate the trend/regression function and the covariance kernel, which can be done using a normal Wishart prior:

$$
\pi(\mu, \Theta) = \mathrm{MVN}(\mu|\Theta)\,\mathrm{Wish}(\Theta)
$$

References

Bergsma, W. (2019). Regression with I-priors. *Econometrics and Statistics*.

Bergsma, W. and Jamil, H. (2023). Additive interaction modelling using I-priors. *arXiv preprint arXiv:2007.15766*.

Genton, M. G. (2001). Classes of kernels for machine learning: a statistics perspective. *Journal of machine learning research*, 2(Dec):299–312.

Yaglom, A. M. (1987). *Correlation Theory of Stationary and Related Random Functions, Volume I: Basic Results*, volume 131. Springer.