Critique of Nested Hierarchical Trans-Dimensional Models

Bill Menke, July 9, 2020

Synopsis. The posterior probability of dimension is not meaningful when the trans-dimensional model parameterization allows a high-dimensional model to be exactly equivalent to a low-dimensional model (in the sense of predicting all the same data).

Introduction

I am concerned here with a trans-dimensional inversion for model parameters $\mathbf{m}$ using observed data $\mathbf{d}^{obs}$ based on Bayes theorem:

$$P(\mathbf{m}|\mathbf{d}^{obs}) = \frac{P(\mathbf{d}^{obs}|\mathbf{m})P_A(\mathbf{m})}{P(\mathbf{d}^{obs})}$$

Here $\mathbf{m}$ is understood to be trans-dimensional and consisting of a set of N-dimensional vectors $\mathbf{m}^{(N)}$ $(N = N^{min} \cdots N^{max})$, each of which is associated with a model space $\mathbb{M}^N$. For simplicity, I restrict model parameters to be on an integers lattice $\mathbb{M}^N = \mathbb{Z}_p^N$; that is, with $\left|m_i^{(N)}\right| \leq p^{(N)}$, where $p^{(N)}$ is a positive integer. The prior distribution is denoted $P_A(\mathbf{m})$. The posterior probability $P(\mathbf{m}|\mathbf{d}^{obs})$ of the model can be written:

$$P(\mathbf{m}|\mathbf{d}^{obs}) = P\left(\mathbf{m}^{(N)}|N, \mathbf{d}^{obs}\right) P(N|\mathbf{d}^{obs})$$

where $P(N|\mathbf{d}^{obs})$ is the posterior probability of the model being N-dimensional. My interest here is the behavior of $P(N|\mathbf{d}^{obs})$ when the model is "hierarchical".

By "hierarchical", I mean the special case where models in the sequence $\mathbf{m}^{(1)}$, $\mathbf{m}^{(2)}$, $\mathbf{m}^{(3)}$ $\cdots$ have increasing complexity. A sequence of Fourier cosine series:

$$m(x) = \sum_{k=0}^{N} m_k^{(N)} \cos(2\pi n x) \quad \text{with} \quad N = 1, 2, 3 \cdots M$$

is of this form, for its ability to represent a complicated $m(x)$ increases with N. A property of this trans-dimensional representations is that an $(N-1)$ dimensional representation is "nested" within an N dimensional one, in the sense that the latter is made equivalent to the former by requiring $m_N^{(N)} = 0$. As I shall argue below, this redundancy in the trans-dimensional model representation produces undesirable behavior in Bayesian estimates of the probability $P(\mathrm{N}|\mathbf{d}^{obs})$

Motivating Example

Consider a trans-dimensional model $\mathbf{m}$ that is either a 2-vector (dimension $N = 2$) or a 3-vector (dimension $N = 3$), where the elements of the vectors take on the binary values (0,1). This simple trans-dimensional model has $2^2 + 2^3 = 4 + 8 = 12$ discrete states.

Suppose that the prior probability $P_A(N=2)$ of the model being a 2-vector is ½ and the model being a 3-vector is ½. Furthermore, suppose that if the model is a 2-vector, then its $J_2 = 4$ states have equal probability $P_A(\mathbf{m}|N=2) = 1/4$ and that if it is a three-vector its $J_3 = 8$ states have equal probability $P_A(\mathbf{m}|N=3) = 1/8$. Here, J_N is the number of model states with dimension N.

Suppose that the observations consist of one datum $d = \sum_i m_i$ and that it is measured with sufficient accuracy that $P(d^{obs}|\mathbf{m}) \approx 1$ when $d^{obs} = d(\mathbf{m})$ and that $P(d^{obs}|\mathbf{m}) \approx 0$ when $d^{obs} \neq d(\mathbf{m})$.

Now suppose $d^{obs} = 0$. We apply Bayes theorem to calculate the posterior probability $P(\mathbf{m}|d^{obs})$:

$$P(\mathbf{m}|d^{obs}) = \frac{P(d^{obs}|\mathbf{m})P_A(\mathbf{m})}{P(d^{obs})} = \frac{P(d^{obs}|\mathbf{m})P_A(\mathbf{m})}{\sum_i P(d^{obs}|\mathbf{m}^{(i)})P(\mathbf{m}^{(i)})}$$

$$\text{with} \sum_i P(\mathrm{d}|\mathbf{m}^{(i)})P(\mathbf{m}^{(i)}) = \frac{1}{8} + \frac{1}{16} = \frac{3}{16}$$

The probability that the dimension is $N = 2$ is:

$$P(N=2|d^{obs}) = \sum_{2-vector\ states\ i} P(\mathbf{m}^{(i)}|d^{obs}) = \left(\frac{1}{8}\right) \Big/ \left(\frac{3}{16}\right) = \frac{2}{3}$$

and the probability that the dimension is $N = 3$ is:

$$P(N=3|d^{obs}) = \sum_{3-vector\ states\ i} P(\mathbf{m}^{(i)}|d^{obs}) = \left(\frac{1}{16}\right) \Big/ \left(\frac{3}{16}\right) = \frac{1}{3}$$

Similar calculations can be performed for $d^{obs} = 1$ and $d^{obs} = 2$ (see Appendix). The results are summarized as follows:

d^{obs}	$P(N=2\|d^{obs})$	$P(N=3\|d^{obs})$
$prior$	1/2	1/2
0	2/3	1/3
1	4/7	3/7
2	2/5	3/5
3	0	1

The probability of the dimension is a strong function of d^{obs}. The $N = 2$ model is most probable when $d^{obs} \leq 1$ and the $N = 3$ model when $d^{obs} = 2$. Superficially, this behavior is acceptable; in general, we expect the data to provide information on the dimension.

But *what* aspect of the data provides information on dimensionality? The data is just the sum of the model parameters. *How* does the process of summing discriminate between possible dimensions?

Analysis

Bayes Theorem indicates that, in the above example, the ratio r of dimensional probabilities is:

$$r = \frac{P(N = 2|d^{obs})}{P(N = 3|d^{obs})} = \frac{\sum_{2-vector\ states\ i} P\left(\mathbf{m}^{(i)}|d^{obs}\right)}{\sum_{3-vector\ states\ i} P\left(\mathbf{m}^{(i)}|d^{obs}\right)} =$$

$$= \frac{\sum_{2-vector\ states\ i} \left\{\begin{matrix} 1 \text{ if } d^{obs} = d\left(\mathbf{m}^{(i)}\right) \\ 0 \text{ otherwise} \end{matrix}\right\} P_A\left(\mathbf{m}^{(i)}\right)}{\sum_{3-vector\ states\ i} \left\{\begin{matrix} 1 \text{ if } d^{obs} = d\left(\mathbf{m}^{(i)}\right) \\ 0 \text{ otherwise} \end{matrix}\right\} P_A\left(\mathbf{m}^{(i)}\right)} =$$

$$= \frac{\sum_{\substack{2-vector\ states\ i \\ that\ satisy\ the\ data}} P_A\left(\mathbf{m}^{(i)}\right)}{\sum_{\substack{3-vector\ states\ i \\ that\ satisy\ the\ data}} P_A\left(\mathbf{m}^{(i)}\right)}$$

In the special case where the prior asserts that all model states with the same dimension have equal probability, $P_A\left(\mathbf{m}^{(i)}|N\right) = 1/J_N$, where J_N is the number of model states in dimension N. If K_N is the number of states of an N-vector that satisfy the data (the "multiplicity"), then:

$$r = \frac{K_2}{J_2} \Big/ \frac{K_3}{J_3}$$

Consequently, the ratio depends on the fraction of model states that satisfy the data for each dimension. I consider this behavior to be "unintuitive". Given that at least one state in each dimension satisfies the data exactly, my intuitive notion is that the ratio of probabilities should be given by the prior. However, since:

$$r \neq \frac{J_3}{J_2}$$

my intuitive notion is wrong! The value of r depends on the number degree of non-uniqueness of the different dimensions, too.

Criticism

My finding that the degree of non-uniqueness affects the posterior probability of dimension is especially troublesome in hierarchical models, when a lower dimensional model is nested within a higher dimensional model, and especially when many distinct instances of the lower dimensional model are nested within the higher dimensional one.

The following layered parameterization is a nested set of hierarchical models: Suppose that the layering is in the x-direction in the interval $(0, X)$. A model with n layers has $N = 2n - 1$ model parameters: $(n - 1)$ layer thicknesses h_i and n layer material properties v_i. Suppose also that the data depends only upon the x-variation of $v(x)$ and not on the presence or absence of layer interfaces. In this case, a model with a small number of layers is multiply nested within the space of models with a larger number of layers. An example with a multiplicity of three is shown (Figure 1). Furthermore, the multiplicity is not the same for all choices of the lower-dimensional model. For instance, the example in Figure 1 would have a multiplicity of four if, for all layers, $v_i = 7$.

When the prior assigns all dimensions equally probability, and all model states within a dimension equal probability; then $P_A\left(\mathbf{m}^{(N)}|N\right) = 1/J_N$ declines rapidly with N. Suppose that the data is exactly satisfied by a state of low dimension N_0. The multiplicity K_N of this state within a higher dimension may grow rapidly with N. Thus, $P(N|d^{obs}) \propto K_N/J_N$ may be peaked at some dimension $N > N_0$. While the position of this peak depends upon the data, it does so only because multiplicity is a data-dependent function; that is $K_N(d^{obs})$. Thus, the posterior probability of the dimension has very little to do with any property of the data, and a lot to do with the combinatoric properties of the model parameterization.

However, I have found that constructing a "natural" non-nested hierarchy of models to be difficult. However, any nested parameterization can be "unnested" by explicitly excluding from $\mathbb{M}^N$ all model states that appear within any lower dimensional space.

The Fourier cosine representation mentioned earlier can be made into a non-nested parameterization by requiring that $m_N^{(N)} \neq 0$. The condition only eliminates a state on the boundary of $\mathbb{Z}_p^N$, and does not open "holes" within it. My sense is that model spaces with complicated topologies should be avoided, because they complicate the application of sampling algorithms such as Metropolis-Hastings.

Another "almost perfect" parameterization is a set of layered models where the number of layers is a prime number, the layers of a given dimension all have equal thickness, and model parameters are layer properties (Figure 2). No higher dimensional model is exactly equivalent to any given lower dimensional one, except for the trivial case of all layer properties being equal. This $1:1:\cdots:1$ line through $\mathbb{Z}^N$ needs to be explicitly excluded from the spaces $\mathbb{M}^N$, $N > 1$ (which creates a "hole" through them).

Parameterizations in which the model space $\mathbb{M}^N$ contains redundancies are problematic even when no nesting occurs, because a "physically-reasonable" prior may be difficult to state. Consider the case where the data depends only upon $v(x)$ and where that function is represented as a Gaussian Process with n training points; that is, $N = 2n$ and $\mathbf{m}^{(N)} = [x_1, v_1, x_2, v_2, \cdots\ x_n, v_n]^{\,T}$. Because the order of training points is arbitrary, a given function $v(x)$ is represented $n!$ times within the model space. The prior assumption that all dimensions are equally probable and all model states within a dimension are equally probable implies $P_A\left(\mathbf{m}^{(N)}\right) = 1/(MJ_N)$. Changing the latter assumption to *distinguishable* states having equal probability implies $P_A\left(\mathbf{m}^{(N)}\right) = (½N)!/(MJ_N)$. The latter formula would seem to be the more "physically-reasonable" definition of the prior. However, stating it requires a detailed understanding of the redundancies inherent in particular parameterization, which may not be available in all cases.

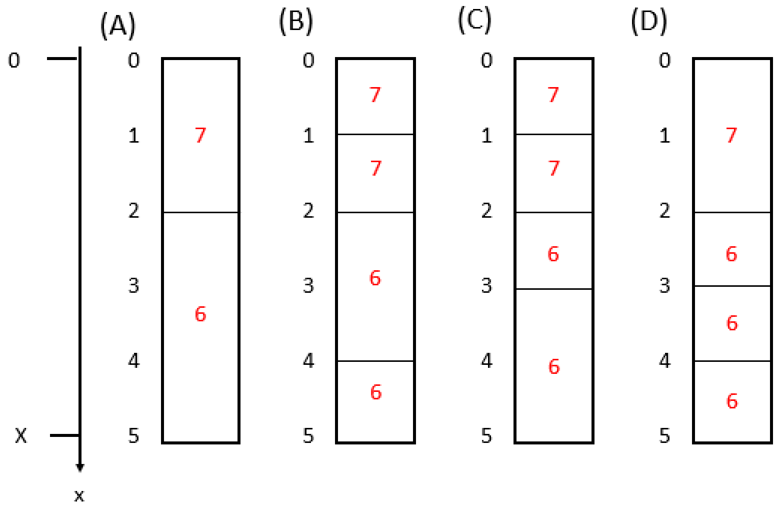

Fig. 1. A nested hierarchical trans-dimensional layered model in which interfaces must be at integer positions x_i (black numerals) with $0 < x < 5$ and layer property v_i (red numerals) must be an integer value $0 < v_i < 11$. The two-layer model (A) is equivalent to three different four-layer models (B-D).

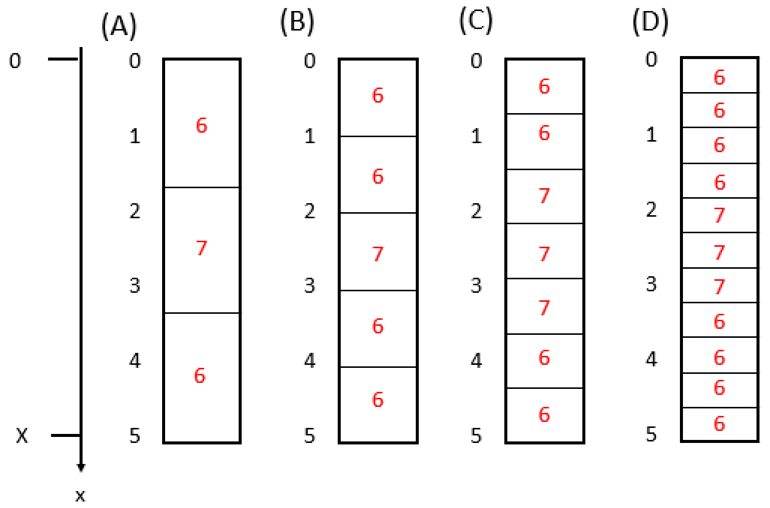

Fig. 2. A non-nested hierarchical trans-dimensional layered model in which interfaces must be evenly distributed between $0 < x_i < 5$ (black numerals) and layer properties v_i (red numerals) must be an integer value $0 < v_i < 11$. No three-layer model (A) is equivalent to any 5, 7 or 11 layer model, though some are close (B-D).

Appendix.

Case of $d^{obs} = 1$. Two 2-vector states $(1{,}0)$ and $(0{,}1)$ and three 3-vector states $(1{,}0{,}0)$ $(0{,}1{,}0)$ and $(0{,}0{,}1)$ have non-zero probability, so:

$$\sum_i P\left(\mathrm{d}|\mathbf{m}^{(i)}\right)P\left(\mathbf{m}^{(i)}\right) = \frac{2}{8} + \frac{3}{16} = \frac{7}{16}$$

The probability that the dimension is $N = 2$ is:

$$P(N = 2|d^{obs}) = \sum_{2-vector\ states\ i} P\left(\mathbf{m}^{(i)}|d^{obs}\right) = \left(\frac{2}{8}\right)\Big/\left(\frac{7}{16}\right) = \frac{4}{7}$$

and the probability that the dimension is $N = 3$ is:

$$P(N = 3|d^{obs}) = \sum_{3-vector\ states\ i} P\left(\mathbf{m}^{(i)}|d^{obs}\right) = \left(\frac{3}{16}\right)\Big/\left(\frac{7}{16}\right) = \frac{3}{7}$$

Case of $d^{obs} = 2$. One 2-vector states $(1{,}1)$ and three 3-vector states $(1{,}1{,}0)$ $(1{,}0{,}1)$ and $(0{,}1{,}1)$ have non-zero probability, so:

$$\sum_i P\left(\mathrm{d}|\mathbf{m}^{(i)}\right)P\left(\mathbf{m}^{(i)}\right) = \frac{1}{8} + \frac{3}{16} = \frac{5}{16}$$

The probability that the dimension is $N = 2$ is:

$$P(N = 2|d^{obs}) = \sum_{2-vector\ states\ i} P\left(\mathbf{m}^{(i)}|d^{obs}\right) = \left(\frac{1}{8}\right)\Big/\left(\frac{5}{16}\right) = \frac{2}{5}$$

and the probability that the dimension is $N = 3$ is:

$$P(N = 3|d^{obs}) = \sum_{3-vector\ states\ i} P\left(\mathbf{m}^{(i)}|d^{obs}\right) = \left(\frac{3}{16}\right)\Big/\left(\frac{5}{16}\right) = \frac{3}{5}$$

Case of $d^{obs} = 3$. No 2-vector state and one 3-vector states $(1{,}1{,}1)$ have non-zero probability, so:

$$\sum_i P\left(\mathrm{d}|\mathbf{m}^{(i)}\right)P\left(\mathbf{m}^{(i)}\right) = 0 + \frac{1}{16} = \frac{1}{16}$$

The probability that the dimension is $N = 2$ is: $P(N = 2|d^{obs}) = 0$ and the probability that the dimension is $N = 3$ is:

$$P(N = 3|d^{obs}) = \sum_{3-vector\ states\ i} P\left(\mathbf{m}^{(i)}|d^{obs}\right) = \left(\frac{1}{16}\right)\Big/\left(\frac{1}{16}\right) = 1$$