Supplementary Information

The probabilistic interpretation of the regularization parameter λ

Ridge regression. Here, we derive the (known) result that if (1) the coefficients β_j are normally distributed with variance τ^2, i.e. $\beta \sim N(0, \tau^2 \mathbf{I})$, and (2) that the training data y_i are contaminated with Gaussian noise of variance σ^2, i.e. $\mathbf{y} \sim N(\mathbf{X}\beta, \sigma^2 \mathbf{I})$, then the problem of finding the maximum likelihood fit is equivalent to regularization by ridge regression, with $\lambda = \sigma^2/\tau^2$.

The log-likelihood of obtaining the training data $\mathbf{y}$ from this model is

$$\log P(\mathbf{y}|\sigma, \tau, \beta) = N \log\left(\frac{1}{\sqrt{\pi\sigma^2}}\right) + K \log\left(\frac{1}{\sqrt{\pi\tau^2}}\right) - \sum_i^N \frac{(y_i - \sum_j X_{ij}\beta_j)^2}{2\sigma^2} - \sum_j^K \frac{\beta_j^2}{2\tau^2}$$

The maximum likelihood $\hat{\beta}$ can be found by applying Bayes' identity

$$P(\beta|\sigma, \tau, \mathbf{y}) \propto \frac{P(\mathbf{y}|\sigma, \tau, \beta)}{P(\mathbf{y})}$$

and setting

$$\frac{\partial \log P(\beta|\sigma, \tau, \mathbf{y})}{\partial \beta} = 0$$

$$\frac{\partial}{\partial \beta}\left(-\sum_i^N \frac{(y_i - \sum_j X_{ij}\beta_j)^2}{2\sigma^2} - \sum_j^K \frac{\beta_j^2}{2\tau^2}\right) = 0$$

$$\frac{\partial}{\partial \beta}\left(\sum_i^N (y_i - \sum_j X_{ij}\beta_j)^2 + \frac{\sigma^2}{\tau^2}\sum_j^K \beta_j^2\right) = 0$$

This is equivalent to finding the β that minimizes $(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)+\lambda\beta^T\beta$, where $\lambda = \sigma^2/\tau^2$.

Lasso regression. If the β_j are distributed as $\beta_j \sim \frac{1}{2\tau^2}\exp(-\frac{|\beta_j|}{\tau^2})$, then the problem of finding the maximum likelihood fit is equivalent to regularization by lasso regression, with $\lambda = \sigma^2/\tau^2$. The log-likelihood of obtaining the training data $\mathbf{y}$ from this model is

$$\log P(\beta|\mathbf{y},\sigma,\tau) = N\log\left(\frac{1}{\sqrt{\pi\sigma^2}}\right) + K\log\left(\frac{1}{2\tau^2}\right) - \sum_i^N \frac{(y_i - \sum_j X_{ij}\beta_j)^2}{2\sigma^2} - \sum_j^K \frac{|\beta_j|}{\tau^2}$$

and the maximum likelihood $\hat{\beta}$ can be found by setting

$$\frac{\partial}{\partial\beta}\left(\sum_i^N \frac{1}{2}(y_i - \sum_j X_{ij}\beta_j)^2 + \frac{\sigma^2}{\tau^2}\sum_j^K |\beta_j|\right) = 0$$

Elastic net regression. If the β_j are distributed as $\beta_j \sim P(\beta_j) = A(\tau,\rho)\exp(-\rho\frac{\beta_j^2}{2\tau^2} - (1-\rho)\frac{|\beta_j|}{\tau^2})$, where ρ is a mixing parameter and $A(\tau,\rho)$ is the normalization constant that makes $\int P(\beta_j)d\beta_j = 1$, then the problem of finding the maximum likelihood fit is equivalent to regularization by elastic net regression, with $\lambda = \sigma^2/\tau^2$. The log-likelihood of obtaining the training data $\mathbf{y}$ from this model is

$$\log P(\beta|\mathbf{y},\sigma,\tau) = N\log\left(\frac{1}{\sqrt{\pi\sigma^2}}\right) + K\log\left(\sqrt{\frac{\pi}{2\rho}}\tau\right) + K\left(\gamma^2 + \log(1+\mathrm{Erf}(\gamma))\right)$$

$$-\sum_i^N \frac{(y_i - \sum_j X_{ij}\beta_j)^2}{2\sigma^2} - \sum_j^K \rho\frac{\beta_j^2}{2\tau^2} - \sum_j^K (1-\rho)\frac{|\beta_j|}{\tau^2}$$

where $\gamma \equiv \frac{\rho-1}{\sqrt{2\rho\tau^2}}$. The maximum likelihood $\hat{\beta}$ in this case can be found by setting

$$\frac{\partial}{\partial\beta}\left(\sum_{i}^{N}\frac{1}{2}(y_i - \sum_j X_{ij}\beta_j)^2 + \frac{\sigma^2}{\tau^2}\sum_j^K |\beta_j| + \frac{\sigma^2}{2\tau^2}\sum_j^K \beta_j^2\right) = 0$$

Standardizing of input data

Here we mention an important property of ridge, lasso and elastic regression: the results of the regression are not invariant to scaling of the training data, i.e. scaling or translating the training data $\mathbf{y}$ will non-trivially influence the effects of the regularization penalty. Thus, in practice, it is a good idea to *standardize* the input data by some method before calculating rate spectra. If the time series has a non-zero baseline (i.e. $y(t)$ does not go to zero as $t \to \infty$), a typical procedure would be to eliminate this degree of freedom by "centering" the data. First, assuming that β_K corresponds to $k_K = 0$, one can estimate β_K directly by $\bar{y} = \frac{1}{N}\sum_i y_i$. Then, for the remaining β_j, $j = 1, ..., K-1$, $\bar{y}$ is subtracted from each y_i, and the regressor inputs X_{ij} are replaced by $X_{ij} - (1/N)\sum_i X_{ij}$. ($\mathbf{X}$ now being a $(N+K-1) \times K-1$ matrix.) This prevents the baseline from being penalized by the regularization procedure.

For the purposes of computing rate spectra, however, the centering procedure is not robust. We find that $\bar{y}$ is not always an accurate estimate of β_K, depending on the form of the input data. Instead, simply scaling the input data y_i to the interval $[0, 1]$, and including β_K in the regression, gives better results. While the baseline is still penalized by regularization, the effect is negligible, especially when spectrum includes many other rates k_j near zero.

Rate spectra for stretched-exponential functions

The analytical solution for the (continuous) rate spectrum $H_\gamma(k)$, is given by two equivalent formulae [10]:

$$H_\gamma(k) = \frac{\tau_0}{\pi}\int_0^\infty \exp(-k\tau_0 u)\exp[-u^\gamma\cos(\gamma\pi)]\sin[u^\gamma\sin(\gamma\pi)]du \tag{9}$$

whose numerical computation works well for large values of k, and

$$H_\gamma(k) = \frac{\tau_0}{\pi}\int_0^\infty \exp[-u^\gamma\cos(\frac{\gamma\pi}{2})]\cos[u^\gamma\sin(\frac{\gamma\pi}{2} - k\tau_0 u)]du \tag{10}$$

which works well for small values of k.

Both derivations proceed from the Bromwich integral

$$H(K) = \lim_{\epsilon\to 0}\frac{1}{2\pi}\int_{\epsilon-i\infty}^{\epsilon+i\infty} I(T)e^{KT}dT$$

where $I(T) = \exp(-T^\beta)$, and $T = t/\tau_0$. Equation (9) proceeds from complex inversion integral by defining a special contour [15], while Equation (10) proceeds from a contour integration [16].

Supplementary Figures

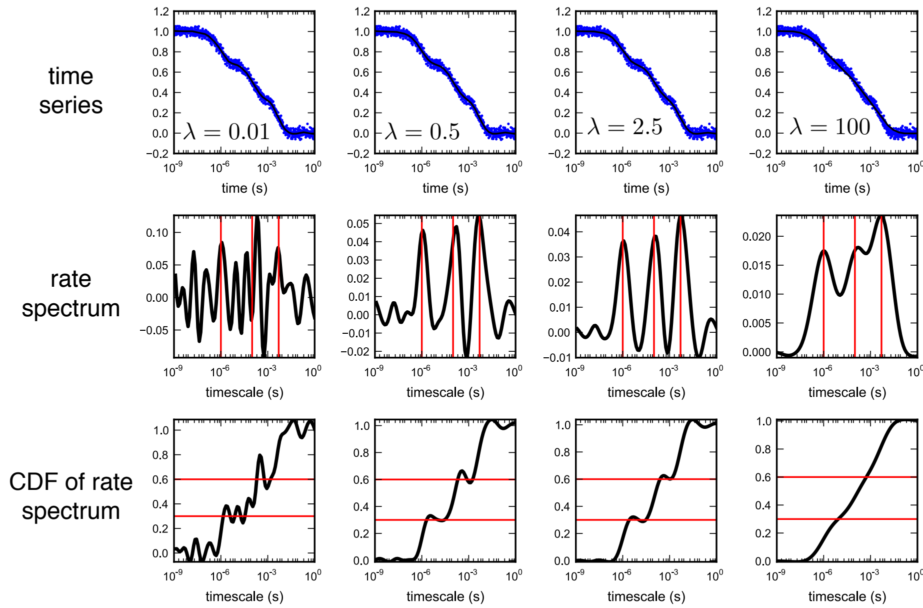

Figure S1: **The effects of regularization (using ridge regression) on a noisy tri-exponential dataset.** Artificial noise $N(0, s^2), s = 0.025$ was added to 1000 samples of a tri-exponential time series with time constants $(\tau_1, \tau_2, \tau_3) = (10^{-6}s, 10^{-4}s, 5 \times 10^{-3}s)$, and amplitudes (respectively) of $(\alpha_1, \alpha_2, \alpha_3) = (0.3, 0.3, 0.4)$. Columns show the results for $\lambda = 0.01, 0.5, 2.5$ and 100: (top) a noisy data set (blue) with best-fit time traces $\hat{\mathbf{y}}$, (middle) the calculated rate spectrum (red lines for each τ_i), and (bottom) the cumulative distribution of rate amplitudes (with red lines indicated the cumulative amplitudes of each relaxation. For small values of λ (0.01), the spectrum is only weakly regularized, resulting in a spectrum heavily affected by noise. For larger values of λ (0.5, 2.5), three peaks corresponding to each timescale in the data is recovered. For very large values of λ (100), the rate spectrum is broadened, although in this case, the three timescales are still discernible.

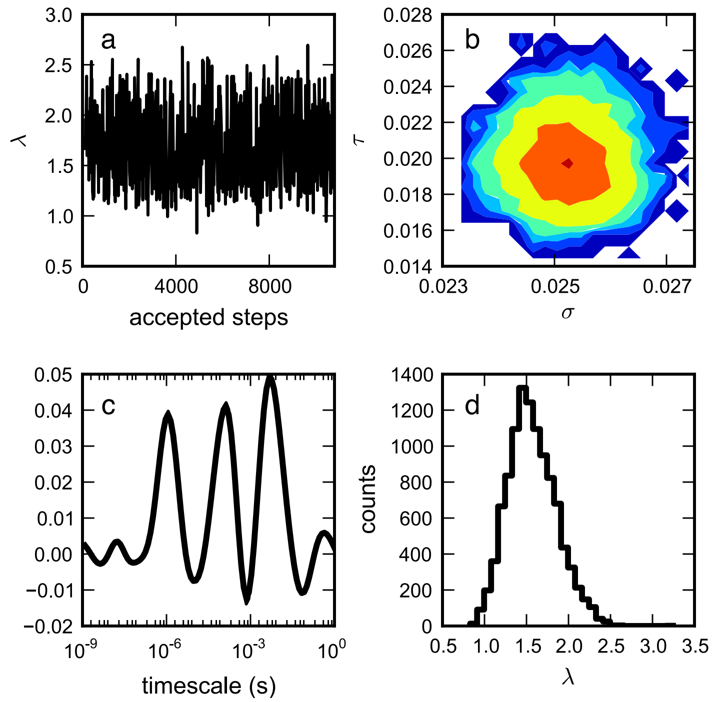

Figure S2: **Posterior sampling of the regularization parameter λ.** Ridge regression was performed for the tri-exponential data with added noise ($s = 0.05$). (a) Monte Carlo sampling of the posterior in σ and τ produces a converged trajectory of values $\lambda = \sigma^2/\tau^2$. (b) A contour plot of (σ, τ) counts (contours from blue to red: 1, 2, 5, 10, 25, 100, 250, 500). (c) The rate spectrum calculated as the expectation over all posterior samples. (d) The posterior distribution of $P(\lambda|\mathbf{y})$, as calculated from posterior sampling.